# 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 2018-19, HKUST, Hong Kong

# Acknowledgment

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


- [Hun-Lan'J04] D. R. Hunter, and K. Lange, "A Tutorial on MM Algorithms", *Amer. Statistician*, pp. 30-37, 2004.
- [Raz-Hon-Luo'J13] M. Razaviyayn, M. Hong, and Z. Luo, "A Unified Convergence Analysis of Block Successive Minimization Methods for Nonsmooth Optimization", *SIAM J. Optim.*, pp. 1126-1153, 2013.
- [Scu-Fac-Son-Pal-Pan'J14] G. Scutari, F. Facchinei, Peiran Song,D. P. Palomar, and Jong-Shi Pang, "Decomposition by Partial Linearization: Parallel Optimization of Multi-Agent Systems", *IEEE Trans. Signal Processig*, vol. 62, no. 3, pp. 641-656, Feb. 2014.
- [Sun-Bab-Pal'J17] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning", *IEEE Trans. Signal Process*, vol. 65, no. 3, pp. 794-816, Feb. 2017.

# Outline

# Outline

# Problem Statement

- Consider the following optimization problem
$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{X}, \end{array}$$
with $\mathcal{X}$ being a closed convex set and $f(\mathbf{x})$ being continuous.
- $f(\mathbf{x})$ is too complicated to manipulate.
- Idea: successively minimize an approximating function $u(\mathbf{x}, \mathbf{x}^k)$
$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} u\left(\mathbf{x}, \mathbf{x}^k\right),$$
hoping the sequence of minimizers $\{\mathbf{x}^k\}$ will converge to optimal $\mathbf{x}^\star$.
- Question: how to construct $u(\mathbf{x}, \mathbf{x}^k)$?

## Problem Statement

- Consider the following optimization problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{X}, \end{array}$$

  with $\mathcal{X}$ being a closed convex set and $f(\mathbf{x})$ being continuous.
- $f(\mathbf{x})$ is too complicated to manipulate.
- Idea: successively minimize an approximating function $u(\mathbf{x}, \mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} u\left(\mathbf{x}, \mathbf{x}^k\right),$$

  hoping the sequence of minimizers $\{\mathbf{x}^k\}$ will converge to optimal $\mathbf{x}^\star$.
- Question: how to construct $u(\mathbf{x}, \mathbf{x}^k)$?

# Problem Statement

- Consider the following optimization problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{X}, \end{array}$$

  with $\mathcal{X}$ being a closed convex set and $f(\mathbf{x})$ being continuous.
- $f(\mathbf{x})$ is too complicated to manipulate.
- Idea: successively minimize an approximating function $u(\mathbf{x}, \mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} u\left(\mathbf{x}, \mathbf{x}^k\right),$$

  hoping the sequence of minimizers $\{\mathbf{x}^k\}$ will converge to optimal $\mathbf{x}^\star$.
- Question: how to construct $u(\mathbf{x}, \mathbf{x}^k)$?

# Terminology

- Distance from a point to a set:

$$d\left(\mathbf{x},\mathscr{S}\right)=\inf_{\mathbf{s}\in\mathscr{S}}\left\|\mathbf{x}-\mathbf{s}\right\|.$$

- Directional derivative:

$$f'\left(\mathbf{x};\mathbf{d}\right)\triangleq\liminf_{\lambda\downarrow 0}\frac{f\left(\mathbf{x}+\lambda\mathbf{d}\right)-f\left(\mathbf{x}\right)}{\lambda}.$$

- Stationary point: $\mathbf{x}$ is a stationary point if

$$f'\left(\mathbf{x};\mathbf{d}\right)\geq 0,\ \forall\mathbf{d}\text{ such that }\mathbf{x}+\mathbf{d}\in\mathscr{X}.$$

# Majorization-Minimization

- Construction rule:

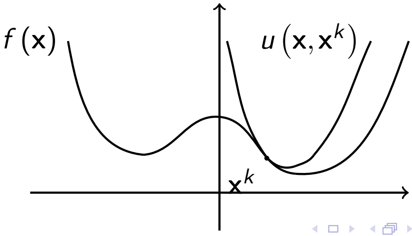
$$u(\mathbf{y},\mathbf{y}) = f(\mathbf{y}),\ \forall \mathbf{y} \in \mathscr{X} \tag{A1}$$

$$u(\mathbf{x},\mathbf{y}) \geq f(\mathbf{x}),\ \forall \mathbf{x},\mathbf{y} \in \mathscr{X} \tag{A2}$$

$$u'(\mathbf{x},\mathbf{y};\mathbf{d})\big|_{\mathbf{x}=\mathbf{y}} = f'(\mathbf{y};\mathbf{d}),\ \forall \mathbf{d} \text{ with } \mathbf{y}+\mathbf{d} \in \mathscr{X} \tag{A3}$$

$$u(\mathbf{x},\mathbf{y}) \text{ is continuous in } \mathbf{x} \text{ and } \mathbf{y} \tag{A4}$$

- Pictorially:



# Algorithm

Majorization-Minimization (Successive Upper-Bound Minimization):

1: Find a feasible point $\mathbf{x}^0 \in \mathscr{X}$ and set $k = 0$
2: **repeat**
3: $\quad \mathbf{x}^{k+1} = \arg\min_{\mathbf{x} \in \mathscr{X}} u\left(\mathbf{x}, \mathbf{x}^k\right)$ (global minimum)
4: $\quad k \leftarrow k + 1$
5: **until** some convergence criterion is met

## Convergence

- Under assumptions A1-A4, every limit point of the sequence $\{\mathbf{x}^k\}$ is a stationary point of the original problem.
- If further assume that the level set $\mathscr{X}^0 = \{\mathbf{x} | f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$ is compact, then

$$\lim_{k \to \infty} d\left(\mathbf{x}^k, \mathscr{X}^\star\right) = 0,$$

where $\mathscr{X}^\star$ is the set of stationary points.

# Outline

## Construct Surrogate Function

- The performance of Majorization-Minimization algorithm depends crucially on the surrogate function $u\left(\mathbf{x},\mathbf{x}^k\right)$.
- Guideline: the global minimizer of $u\left(\mathbf{x},\mathbf{x}^k\right)$ should be easy to find.
- Suppose $f\left(\mathbf{x}\right)=f_1\left(\mathbf{x}\right)+\kappa\left(\mathbf{x}\right)$, where $f_1\left(\mathbf{x}\right)$ is some "nice" function and $\kappa\left(\mathbf{x}\right)$ is the one needed to be approximated.

# Construction by Convexity

- 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:

$$\begin{aligned}
\kappa\left(\mathbf{w}^T\mathbf{x}\right) &= \kappa\left(\mathbf{w}^T\left(\mathbf{x}-\mathbf{x}^k\right)+\mathbf{w}^T\mathbf{x}^k\right)\\
&= \kappa\left(\sum_i \alpha_i\left(\frac{w_i\left(x_i-x_i^k\right)}{\alpha_i}+\mathbf{w}^T\mathbf{x}^k\right)\right)\\
&\leq \sum_i \alpha_i \kappa\left(\frac{w_i\left(x_i-x_i^k\right)}{\alpha_i}+\mathbf{w}^T\mathbf{x}^k\right)
\end{aligned}$$

- If further assume that $\mathbf{w}$ and $\mathbf{x}$ are positive ($\alpha_i = w_i x_i^k/\mathbf{w}^T\mathbf{x}^k$):

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

- The surrogate functions are separable (parallel algorithm).

# Construction by Taylor Expansion

- Suppose $\kappa(\mathsf{x})$ is concave and differentiable, then

$$\kappa(\mathsf{x}) \leq \kappa\left(\mathsf{x}^k\right) + \nabla\kappa\left(\mathsf{x}^k\right)\left(\mathsf{x} - \mathsf{x}^k\right),$$

  which is a linear upper-bound.
- Suppose $\kappa(\mathsf{x})$ is convex and twice differentiable, then

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

  if $\mathbf{M} - \nabla^2\kappa(\mathsf{x}) \succeq \mathbf{0}$, $\forall \mathsf{x}$.

# 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:

$$\|\mathbf{x}\| \geq \frac{\mathbf{x}^T\mathbf{x}^k}{\|\mathbf{x}^k\|}$$

- Jensen's Inequality:

$$\kappa(\mathrm{E}\mathbf{x}) \leq \mathrm{E}\kappa(\mathbf{x})$$

with $\kappa(\cdot)$ being convex.

# Outline

# EM Algorithm

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

$$u\left(\theta, \theta^k\right) = -\mathrm{E}_{\mathbf{z}|\mathbf{x},\theta^k} \log p(\mathbf{x}, \mathbf{z}|\theta)$$

    update $\theta$ from "guessed" complete data set

# An MM Interpretation of EM

- The objective function can be written as

$$
\begin{aligned}
& -\log p(\mathbf{x}|\theta) \\
= & -\log \mathrm{E}_{\mathbf{z}|\theta} p(\mathbf{x}|\mathbf{z},\theta) \\
= & -\log \mathrm{E}_{\mathbf{z}|\theta} \left( \frac{p\left(\mathbf{z}|\mathbf{x},\theta^k\right) p(\mathbf{x}|\mathbf{z},\theta)}{p\left(\mathbf{z}|\mathbf{x},\theta^k\right)} \right) \\
= & -\log \mathrm{E}_{\mathbf{z}|\mathbf{x},\theta^k} \left( \frac{p(\mathbf{x}|\mathbf{z},\theta)}{p\left(\mathbf{z}|\mathbf{x},\theta^k\right)} p(\mathbf{z}|\theta) \right) \\
\leq & -\mathrm{E}_{\mathbf{z}|\mathbf{x},\theta^k} \log \left( \frac{p(\mathbf{x}|\mathbf{z},\theta)}{p\left(\mathbf{z}|\mathbf{x},\theta^k\right)} p(\mathbf{z}|\theta) \right) \qquad \text{(Jensen's Inequality)} \\
= & \underbrace{-\mathrm{E}_{\mathbf{z}|\mathbf{x},\theta^k} \log p(\mathbf{x},\mathbf{z}|\theta)}_{u\left(\theta,\theta^k\right)} + \mathrm{E}_{\mathbf{z}|\mathbf{x},\theta^k} p\left(\mathbf{z}|\mathbf{x},\theta^k\right)
\end{aligned}
$$

# Proximal Minimization

- $f(\mathbf{x})$ is convex. Solve $\min_{\mathbf{x}} f(\mathbf{x})$ by solving the equivalent problem

$$\underset{\mathbf{x}\in\mathscr{X},\mathbf{y}\in\mathscr{X}}{\text{minimize}} \quad f(\mathbf{x}) + \frac{1}{2c}\|\mathbf{x}-\mathbf{y}\|^2 .$$

- Objective function is strongly convex in both $\mathbf{x}$ and $\mathbf{y}$.
- Algorithm:

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

- An MM interpretation:

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

# DC Programming

- Consider the unconstrained problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \quad f(\mathbf{x}),$$

where $f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$ with $g(\mathbf{x})$ convex and $h(\mathbf{x})$ concave.

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

$$\nabla g\left(\mathbf{x}^{k+1}\right) = -\nabla h\left(\mathbf{x}^k\right).$$

- An MM interpretation:

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}} \left\{ g(\mathbf{x}) + \nabla h\left(\mathbf{x}^k\right)^T \left(\mathbf{x} - \mathbf{x}^k\right) \right\}.$$

# Outline

# Power Control by GP

- [Chi-Tan-Pal-O'Ne-Jul'J07] Problem: maximize system throughput. Essentially we need to solve the following problem:

$$\underset{\mathbf{P}\in\mathscr{P}}{\text{minimize}} \quad \frac{\sum_{j\neq i} G_{ij}P_j+n_i}{\sum_j G_{ij}P_j+n_i}$$

- Objective function is the ratio of two posynomials.
- Minorize a posynomial, denoted by $g(\mathbf{x}) = \sum_i m_i(\mathbf{x})$, by mononial:

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

where $\alpha_i = \frac{m_i(\mathbf{x}^k)}{g(\mathbf{x}^k)}$. (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

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{x}\|_0 \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \end{array}$$

- $\ell_1$-norm approximation

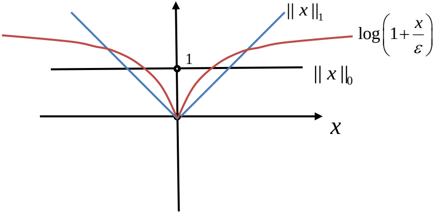
$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{x}\|_1 \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \end{array}$$

- General form

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \sum_{i=1}^{n} \phi\left(|x_i|\right) \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \end{array}$$

- [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

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \sum w_i^k |x_i| \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \end{array}$$



# Sparse Generalized Eigenvalue Problem

- $\ell_0$-norm regularized generalized eigenvalue problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{maximize}} & \mathbf{x}^T\mathbf{A}\mathbf{x} - \rho\,\|\mathbf{x}\|_0 \\ \text{subject to} & \mathbf{x}^T\mathbf{B}\mathbf{x} = 1. \end{array}$$

- 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'J15a] 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

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

- Solve at each iteration the following GEVP:

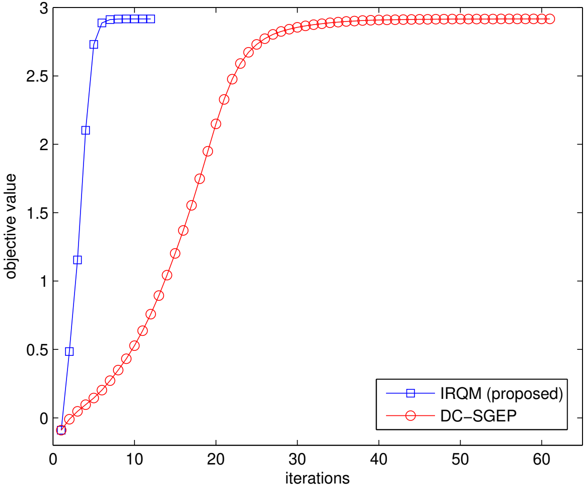
$$\begin{array}{ll} \underset{\mathbf{x}}{\text{maximize}} & \mathbf{x}^T\mathbf{A}\mathbf{x} - \rho\mathbf{x}^T\text{diag}\left(\mathbf{w}^k\right)\mathbf{x} \\ \text{subject to} & \mathbf{x}^T\mathbf{B}\mathbf{x} = 1 \end{array}$$

- 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.


objective value
iterations
IRQM (proposed)
DC–SGEP
0
0.5
1
1.5
2
2.5
3
0
10
20
30
40
50
60

## 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):

$$\mathrm{ISL} = \sum_{k=1}^{N-1} |r_k|^2.$$

- Problem formulation:

$$\begin{array}{ll} \underset{\{x_n\}_{n=1}^N}{\text{minimize}} & \mathrm{ISL} \\ \text{subject to} & |x_n| = 1,\ n = 1, \ldots, N. \end{array}$$

- By Fourier transform:

$$\mathrm{ISL} \propto \sum_{p=1}^{2N} \left[ \left| \mathbf{a}_p^H \mathbf{x} \right|^2 - N \right]^2$$

with $\mathbf{x} = [x_1, \ldots, x_N]^T$, $\mathbf{a}_p = \left[1, e^{j\omega_p}, \ldots, e^{j\omega_p(N-1)}\right]^T$ and $\omega_p = \frac{2\pi}{2N}(p-1)$.

- Equivalent problem:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \sum_{p=1}^{2N} \left( \mathbf{a}_p^H \mathbf{x}\mathbf{x}^H \mathbf{a}_p \right)^2 \\ \text{subject to} & |x_n| = 1, \ \forall n. \end{array}$$

- [Son-Bab-Pal'J15b] Define $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_{2N}]$, $\mathbf{p}^k = \left[\left|\mathbf{a}_1^H \mathbf{x}^k\right|^2, \ldots, \left|\mathbf{a}_{2N}^H \mathbf{x}^k\right|^2\right]^T$, $\tilde{\mathbf{A}} = \mathbf{A}\left(\operatorname{diag}\left(\mathbf{p}^k\right) - p_{\max}^k \mathbf{I}\right)\mathbf{A}^H$.
- Quadratic surrogate function:

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

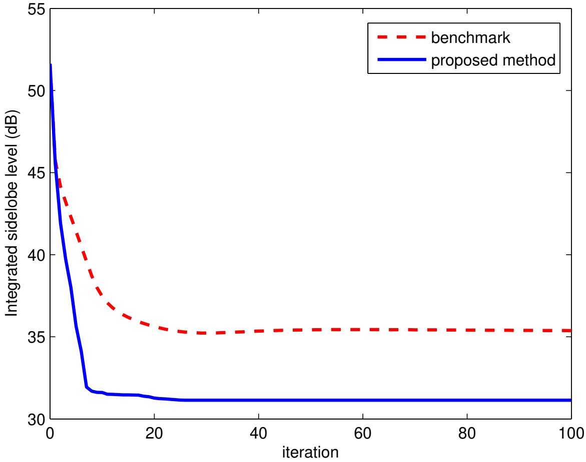
- Equivalent to

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \|\mathbf{x} - \mathbf{y}\|_2 \\ \text{subject to} & |x_n| = 1, \ \forall n \end{array}$$

  with

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

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


55
50
45
40
35
30
0
20
40
60
80
100
Integrated sidelobe level (dB)
iteration
benchmark
proposed method

# Covariance Estimation

- $\mathbf{x}_i \sim \text{elliptical}\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$
- Fitting normalized sample $\mathbf{s}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2}$ to Angular Central Gaussian distribution

$$f\left(\mathbf{s}_i\right) \propto \det\left(\boldsymbol{\Sigma}\right)^{-1/2}\left(\mathbf{s}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{s}_i\right)^{-K/2}$$

- [Sun-Bab-Pal'J14] Shrinkage penalty

$$h\left(\boldsymbol{\Sigma}\right) = \log\det\left(\boldsymbol{\Sigma}\right) + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)$$

- Solve the following problem:

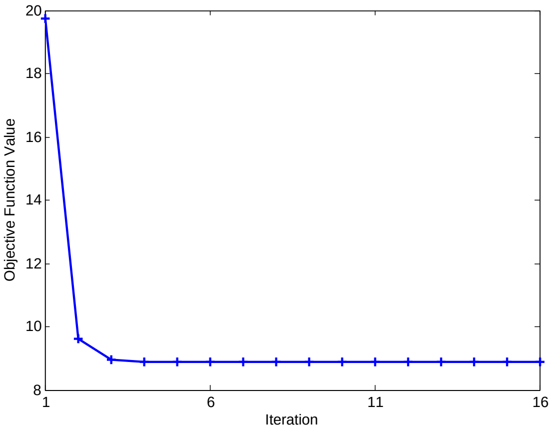
$$\begin{array}{ll} \underset{\boldsymbol{\Sigma}}{\text{minimize}} & \log\det\left(\boldsymbol{\Sigma}\right) + \frac{K}{N}\sum \log\left(\mathbf{x}_i^T \boldsymbol{\Sigma}^{-1}\mathbf{x}_i\right) + \alpha h\left(\boldsymbol{\Sigma}\right) \\ \text{subject to} & \boldsymbol{\Sigma} \succeq \mathbf{0} \end{array}$$

- At $\boldsymbol{\Sigma}^k$, the objective function is majorized by

$$(1+\alpha)\log\det(\boldsymbol{\Sigma})+\frac{K}{N}\sum_{i=1}^{N}\frac{\mathbf{x}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_i}{\mathbf{x}_i^T\left(\boldsymbol{\Sigma}^k\right)^{-1}\mathbf{x}_i}+\alpha\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{T}\right)$$

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

$$\boldsymbol{\Sigma}^{k+1}=\frac{1}{1+\alpha}\frac{K}{N}\sum\frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\left(\boldsymbol{\Sigma}^k\right)^{-1}\mathbf{x}_i}+\frac{\alpha}{1+\alpha}\mathbf{T}$$


20
18
16
14
12
10
8
Objective Function Value
1
6
11
16
Iteration

# Outline

# Problem Statement

- Consider the following problem

$$\underset{\mathbf{x}\in\mathscr{X}}{\text{minimize}} \quad f(\mathbf{x})$$

- Set $\mathscr{X}$ possesses Cartesian product structure $\mathscr{X} = \prod_{i=1}^{m} \mathscr{X}_i$.
- Observation: the problem

$$\underset{\mathbf{x}_i\in\mathscr{X}_i}{\text{minimize}} \quad f\left(\mathbf{x}_1^0,\ldots,\mathbf{x}_{i-1}^0,\mathbf{x}_i,\mathbf{x}_{i+1}^0,\ldots,\mathbf{x}_m^0\right)$$

with $\mathbf{x}_{-i}^0$ taking some feasible value, is easy to solve.

# Outline

# Block Coordinate Descent (BCD)

- Denote $\mathbf{x} \triangleq (\mathbf{x}_1, \ldots, \mathbf{x}_m)$,
  $f\left(\mathbf{x}_1^0, \ldots, \mathbf{x}_{i-1}^0, \mathbf{x}_i, \mathbf{x}_{i+1}^0, \ldots, \mathbf{x}_m^0\right) \triangleq f\left(\mathbf{x}_i, \mathbf{x}^0\right)$
- Block Coordinate Descent (nonlinear Gauss-Seidel)

1: Initialize $\mathbf{x}^0 \in \mathscr{X}$ and set $k = 0$.
2: **repeat**
3: $\quad k = k + 1,\ i = (k \mod n) + 1$
4: $\quad \mathbf{x}_i^k = \arg\min_{\mathbf{x}_i \in \mathscr{X}_i} f\left(\mathbf{x}_i, \mathbf{x}^{k-1}\right)$
5: $\quad \mathbf{x}_i^k \leftarrow \mathbf{x}_i^{k-1},\ \forall k \neq i$
6: **until** some convergence criterion is met

# Convergence

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

  Then every limit point of the sequence $\left\{\mathbf{x}^k\right\}$ 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 Succesive Majorization-Minimization (BS-MM):

1: Initialize $\mathbf{x}^0 \in \mathscr{X}$ and set $k = 0$.
2: **repeat**
3: $\quad k = k+1,\ i = (k \bmod n)+1$
4: $\quad \mathscr{X}^k = \arg\min_{\mathbf{x}_i \in \mathscr{X}_i} u_i\left(\mathbf{x}_i, \mathbf{x}^{k-1}\right)$
5: $\quad$ Set $\mathbf{x}_i^k$ to be an arbitrary element in $\mathscr{X}^k$
6: $\quad \mathbf{x}_i^k \leftarrow \mathbf{x}_i^{k-1},\ \forall k \neq i$
7: **until** some convergence criterion is met

- Generalization of BCD

# Convergence

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

$$u_i(\mathbf{y}_i,\mathbf{y}) = f(\mathbf{y}),\ \forall \mathbf{y} \in \mathscr{X}, \forall i \tag{B1}$$

$$u_i(\mathbf{x}_i,\mathbf{y}) \geq f(\mathbf{y}_1,\ldots,\mathbf{y}_{i-1},\mathbf{x}_i,\mathbf{y}_{i+1},\ldots,\mathbf{y}_n), \quad \forall \mathbf{x}_i \in \mathscr{X}_i, \forall \mathbf{y} \in \mathscr{X}, \forall i \tag{B2}$$

$$u_i'(\mathbf{x}_i,\mathbf{y};\mathbf{d}_i)\big|_{\mathbf{x}_i=\mathbf{y}_i} = f'(\mathbf{y};\mathbf{d}), \quad \forall \mathbf{d} = (\mathbf{0},\ldots,\mathbf{d}_i,\ldots,\mathbf{0}) \text{ such that } \mathbf{y}_i + \mathbf{d}_i \in \mathscr{X}_i, \forall i \tag{B3}$$

$$u_i(\mathbf{x}_i,\mathbf{y}) \text{ is continuous in } (\mathbf{x}_i,\mathbf{y}),\ \forall i \tag{B4}$$

- In short, $u_i(\mathbf{x}_i,\mathbf{x}^k)$ majorizes $f(\mathbf{x})$ on the $i$th block.

- Under assumptions B1-B4, for simplicity additionally assume that $f$ is continuously differentiable,
  - $u_i(\mathbf{x}_i, \mathbf{y})$ is quasi-convex in $\mathbf{x}_i$,
    each subproblem $\min_{\mathbf{x}_i \in \mathscr{X}_i} u_i\left(\mathbf{x}_i, \mathbf{x}^{k-1}\right)$ has a unique solution for any $\mathbf{x}^{k-1} \in \mathscr{X}$,
    then every limit point of $\left\{\mathbf{x}^k\right\}$ is a stationary point.
  - level set $\mathscr{X}^0 = \left\{\mathbf{x} | f(\mathbf{x}) \leq f\left(\mathbf{x}^0\right)\right\}$ is compact,
    each subproblem $\min_{\mathbf{x}_i \in \mathscr{X}_i} u_i\left(\mathbf{x}_i, \mathbf{x}^{k-1}\right)$ has a unique solution for any $\mathbf{x}^{k-1} \in \mathscr{X}$ for at least $m-1$ blocks,
    then $\lim_{k \to \infty} d\left(\mathbf{x}^k, \mathscr{X}^\star\right) = 0$.
- More restrictive assumption than MM due to the cyclic update behavior.

# Outline

# Alternating Proximal Minimization

- Consider the problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f\left(\mathbf{x}_1, \ldots, \mathbf{x}_m\right) \\ \text{subject to} & \mathbf{x}_i \in \mathscr{X}_i, \end{array}$$

  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

$$\begin{array}{ll} \underset{\mathbf{x}_i}{\text{minimize}} & f\left(\mathbf{x}_1^k, \ldots, \mathbf{x}_{i-1}^k, \mathbf{x}_i, \mathbf{x}_{i+1}^k, \ldots, \mathbf{x}_m^k\right) + \frac{1}{2c}\left\|\mathbf{x}_i - \mathbf{x}_i^k\right\|^2 \\ \text{subject to} & \mathbf{x}_i \in \mathscr{X}_i \end{array}$$

- Strictly convex objective$\rightarrow$unique minimizer

# Proximal Splitting Algorithm

- Consider the following problem

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \sum_{i=1}^{m} f_i(\mathbf{x}_i) + f_{m+1}(\mathbf{x}_1, \ldots, \mathbf{x}_m) \\ \text{subject to} & \mathbf{x}_i \in \mathscr{X}_i, \; i = 1, \ldots, m \end{array}$$

with $f_i$ convex and lower semicontinuous, $f_{m+1}$ convex and

$$\|\nabla f_{m+1}(\mathbf{x}) - \nabla f_{m+1}(\mathbf{y})\| \leq \beta_i \|\mathbf{x}_i - \mathbf{y}_i\|$$

- Cyclically update:

$$\mathbf{x}_i^{k+1} = \text{prox}_{\gamma f_i}\left(\mathbf{x}_i^k - \gamma \nabla_{\mathbf{x}_i} f_{m+1}\left(\mathbf{x}^k\right)\right),$$

with the proximity operator defined as

$$\text{prox}_f(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathscr{X}} f(\mathbf{y}) + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2.$$

# Proximal Splitting Algorithm

- BS-MM interpretation:

$$u_i\left(\mathbf{x}_i,\mathbf{x}^k\right)=f_i\left(\mathbf{x}_i\right)+\frac{1}{2\gamma}\left\|\mathbf{x}_i-\mathbf{x}_i^k\right\|^2+\nabla_{\mathbf{x}_i}f_{m+1}\left(\mathbf{x}^k\right)^T\left(\mathbf{x}_i-\mathbf{x}_i^k\right)$$
$$+\sum_{j\neq i}f_j\left(\mathbf{x}_j^k\right)+f_{m+1}\left(\mathbf{x}_{-i}^k,\mathbf{x}_i\right).$$

- Check:

$$\begin{aligned}&f_{m+1}\left(\mathbf{x}^k\right)+\frac{1}{2\gamma}\left\|\mathbf{x}_i-\mathbf{x}_i^k\right\|^2+\nabla_{\mathbf{x}_i}f_{m+1}\left(\mathbf{x}^k\right)^T\left(\mathbf{x}_i-\mathbf{x}_i^k\right)\\ \geq\, &f_{m+1}\left(\mathbf{x}^k\right)+\frac{\beta_i}{2}\left\|\mathbf{x}_i-\mathbf{x}_i^k\right\|^2+\nabla_{\mathbf{x}_i}f_{m+1}\left(\mathbf{x}^k\right)^T\left(\mathbf{x}_i-\mathbf{x}_i^k\right)\\ \geq\, &f_{m+1}\left(\mathbf{x}_{-i}^k,\mathbf{x}_i\right) \qquad \text{(Descent lemma)}\end{aligned}$$

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

# Outline

# Robust Estimation of Location and Scatter

- $\mathbf{x}_i \sim \text{elliptical}(\mu, \mathbf{R})$
- [Sun-Bab-Pal'J15] Fitting $\mathbf{x}_i$ to a Cauchy distribution with pdf

$$f(\mathbf{x}) \propto \det(\mathbf{R})^{-1/2} \left(1 + (\mathbf{x}_i - \mu)^T \mathbf{R}^{-1} (\mathbf{x}_i - \mu)\right)^{-(K+1)/2}$$

- Shrinkage penalty

$$h(\mathbf{t}, \mathbf{T}) = K \log\left(\mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{T}\right)\right) + \log\det(\mathbf{R}) + \log\left(1 + (\mathbf{t} - \mu)^T \mathbf{R}^{-1} (\mathbf{t} - \mu)\right)$$

- Solve the following problem:

$$\begin{array}{ll} \underset{\mu, \mathbf{R} \succeq \mathbf{0}}{\text{minimize}} & \log\det(\mathbf{R}) + \frac{K+1}{N} \sum_{i=1}^{N} \log\left(1 + (\mathbf{x}_i - \mu)^T \mathbf{R}^{-1} (\mathbf{x}_i - \mu)\right) \\ & + \alpha h(\mathbf{t}, \mathbf{T}) \end{array}$$

- BS-MM Algorithm update:

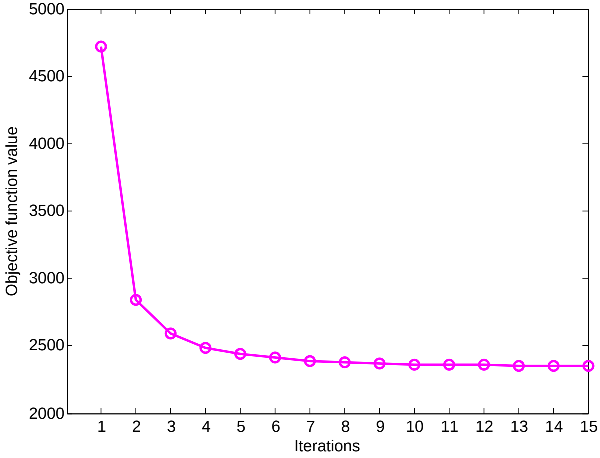
$$\mu_{t+1} = \frac{(K+1)\sum_{i=1}^{N} w_i(\mu_t, \mathbf{R}_t)\mathbf{x}_i + N\alpha w_{\mathbf{t}}(\mu_t, \mathbf{R}_t)\mathbf{t}}{(K+1)\sum_{i=1}^{N} w_i(\mu_t, \mathbf{R}_t) + N\alpha w_{\mathbf{t}}(\mu_t, \mathbf{R}_t)}$$

$$\begin{aligned}\mathbf{R}_{t+1} = {} & \frac{K+1}{N+N\alpha}\sum_{i=1}^{N} w_i(\mu_{t+1}, \mathbf{R}_t)(\mathbf{x}_i - \mu_{t+1})(\mathbf{x}_i - \mu_{t+1})^T \\ & + \frac{N\alpha}{N+N\alpha} w_{\mathbf{t}}(\mu_{t+1}, \mathbf{R}_t)(\mu_{t+1} - \mathbf{t})(\mu_{t+1} - \mathbf{t})^T \\ & + \frac{N\alpha K}{N+N\alpha}\frac{\mathbf{T}}{\mathrm{Tr}\left(\mathbf{R}_t^{-1}\mathbf{T}\right)}\end{aligned}$$

where

$$w_i(\mu, \mathbf{R}) = \frac{1}{1 + (\mathbf{x}_i - \mu)^T \mathbf{R}^{-1}(\mathbf{x}_i - \mu)}$$

$$w_{\mathbf{t}}(\mu, \mathbf{R}) = \frac{1}{1 + (\mathbf{t} - \mu)^T \mathbf{R}^{-1}(\mathbf{t} - \mu)}.$$


5000
4500
4000
3500
3000
2500
2000
Objective function value
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Iterations

# Outline

# Problem Statement

- Consider the following problem:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x}_i \in \mathscr{X}_i \end{array}$$

  where the $\mathscr{X}_i$'s are closed and convex sets, $f(\mathbf{x}) = \sum_{l=1}^{L} f_l(\mathbf{x}_1, \ldots, \mathbf{x}_m)$.
- Conditional gradient update (Frank-Wolfe):

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \gamma^k \mathbf{d}^k$$

  - direction $\mathbf{d}^k \triangleq \bar{\mathbf{x}}^k - \mathbf{x}^k$ with

$$\bar{\mathbf{x}}_i^k = \arg \min_{\mathbf{x}_i \in \mathscr{X}_i} \nabla_{\mathbf{x}_i} f\left(\mathbf{x}^k\right)^T \left(\mathbf{x}_i - \mathbf{x}_i^k\right)$$

  - step-size $\gamma^k \in (0, 1]$, chosen to guarantee convergence.

# Exact Jacobi SCA Algorithm

- Idea:
  - Conditional gradient update linearize all the $f_l$'s at $\mathbf{x}^k$.
  - Each function $f_l$ might be convex w.r.t. some block $\mathbf{x}_i$.
  - We want to preserve the convex property of $f_l\left(\mathbf{x}_i, \mathbf{x}_{-i}^k\right)$.
- Solution: keep the convex $f_l\left(\mathbf{x}_i, \mathbf{x}_{-i}^k\right)$'s and linearize the others.

- Define $\mathscr{C}_i$ as the set of indices of $l$ such that $f_l\left(\mathbf{x}_i, \mathbf{x}_{-i}^k\right)$ is convex.
- Approximate $f\left(\mathbf{x}\right)$ on the $i$th block at point $\mathbf{x}^k$:

$$\tilde{f}_i\left(\mathbf{x}_i, \mathbf{x}^k\right) = \underbrace{\sum_{l \in \mathscr{C}_i} f_l\left(\mathbf{x}_i, \mathbf{x}_{-i}^k\right)}_{\text{convex terms}} + \underbrace{\pi_i\left(\mathbf{x}^k\right)^T\left(\mathbf{x}_i - \mathbf{x}_i^k\right)}_{\substack{\text{linear approx.} \\ \text{non-convex terms}}} + \underbrace{\frac{\tau_i}{2}\left(\mathbf{x}_i - \mathbf{x}_i^k\right)^T \mathbf{H}_i\left(\mathbf{x}^k\right)\left(\mathbf{x}_i - \mathbf{x}_i^k\right)}_{\text{proximal term}},$$

with

$$\pi_i\left(\mathbf{x}^k\right) = \sum_{l \notin \mathscr{C}_i} \nabla_{\mathbf{x}_i} f_l\left(\mathbf{x}^k\right) \text{ and } \mathbf{H}_i\left(\mathbf{x}^k\right) \succ c_{H_i}\mathbf{I}.$$

- Exact Jacobi SCA update:

$$\hat{\mathbf{x}}_i\left(\mathbf{x}^k, \tau_i\right) = \arg\min_{\mathbf{x}_i \in \mathscr{X}_i} \tilde{f}_i\left(\mathbf{x}_i, \mathbf{x}^k\right)$$

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

- 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

# Extensions

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

# Comparison

| | BS-MM | FLEXA |
|---|---|---|
| convergence | stationary point | stationary point |
| objective function | continuous<br>may not be smooth | continuous<br>may not be smooth |
| constraint set | Cartesian | Cartesian & convex |
| update rule | sequential | sequential or parallel |
| approx. function | global upper-bound<br>unique minimizer<br>can be non-convex | local approximation<br>not required<br>convex approx. |

## 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

# References I


D. R. Hunter and K. Lange.
A tutorial on MM algorithms.
*Amer. Statistician*, volume 58, pp. 30–37, 2004.

M. Razaviyayn, M. Hong, and Z. Luo.
A unified convergence analysis of block successive minimization methods for nonsmooth optimization.
*SIAM J. Optim.*, volume 23, no. 2, pp. 1126–1153, 2013.

G. Scutari, F. Facchinei, P. Song, D. Palomar, and J.-S. Pang.
Decomposition by partial linearization: Parallel optimization of multi-agent systems.
*IEEE Trans. Signal Process.*, volume 62, no. 3, pp. 641–656, 2014.

Y. Sun, P. Babu, and D. Palomar.
Majorization-minimization algorithms in signal processing, communications, and machine learning.
*IEEE Transactions on Signal Processing*, volume 65, no. 3, pp. 794–816, 2017.

M. Chiang, C. W. Tan, D. Palomar, D. O'Neill, and D. Julian.
Power control by geometric programming.
*IEEE Trans. Wireless Commun*, volume 6, no. 7, pp. 2640–2651, 2007.

E. J. Candes, M. Wakin, and S. Boyd.
Enhancing sparsity by reweighted l1 minimization.
*J. Fourier Anal. Appl.*, volume 14, no. 5-6, pp. 877–905, 2008.

J. Song, P. Babu, and D. P. Palomar.
Sparse generalized eigenvalue problem via smooth optimization.
*IEEE Trans. Signal Process.*, volume 63, no. 7, pp. 1627–1642, 2015.

—.
Optimization methods for designing sequences with low autocorrelation sidelobes.
*IEEE Trans. Signal Process.*, volume 63, no. 15, pp. 3998–4009, 2015.

Y. Sun, P. Babu, and D. P. Palomar.
Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms.
*IEEE Trans. Signal Processing*, volume 62, no. 19, pp. 5143–5156, 2014.

D. P. Bertsekas.
*Nonlinear Programming*.
Athena Scientific, 1999.

L. Grippo and M. Sciandrone.
On the convergence of the block nonlinear gauss–seidel method under convex constraints.
*Oper. Res. Lett.*, volume 26, no. 3, pp. 127–136, 2000.

Y. Sun, P. Babu, and D. P. Palomar.
Regularized robust estimation of mean and covariance matrix under heavy-tailed distributions.
*IEEE Transactions on Signal Processing*, volume 63, no. 12, pp. 3096–3109, 2015.

# Thanks

For more information visit:

http://www.danielppalomar.com