Some Recent Advances in Alternating Direction Methods: Practice and Theory

Yin Zhang (张寅)

Department of Computational and Applied Mathematics
Rice University, Houston, Texas, USA

The 5th Sino-Japanese Optimization Meeting
Beijing, China, September 28, 2011

Outline:

- Alternating Direction Method (ADM)
- Recent Revival and Extensions
- Local Convergence and Rate
- Global Convergence
- Summary

Contributors:

Xin Liu, Junfeng Yang, Zaiwen Wen, Yilun Wang, Chengbo Li, Yuan Shen, Wei Deng, Wotao Yin

Basic Ideas

To an extent, constructing algorithm $\approx$ "Art of Balance"

- Optimization algorithms are "always" iterative
- Total cost = (number of iterations)×(cost/iter)
- 2 objectives above

It's more difficult to analyze iteration complexity.
A good iteration complexity $\neq$ fast algorithm

ADM Idea: lower per-iteration complexity

Approach:

— "远交近攻", "各个击破" — Sun-Tzu (400 BC)
— "Divide and Conquer" — Julius Caesar (100-44 BC)

RICE

Convex program with the 2-separability structure

$$\min_{x,y} \underbrace{f_1(x) + f_2(y)}_{f(x,y)}, \text{ s.t. } Ax + By = b, \; x \in \mathcal{X}, \; y \in \mathcal{Y}$$

Augmented Lagrangian (AL): penalty $\beta > 0$

$$\mathcal{L}_{\mathcal{A}}(x, y, \lambda) \;\; = \;\; f(x, y) - \lambda^\top (Ax + By - b) + \frac{\beta}{2}\|Ax + By - b\|^2$$

Classic AL Multipler Method (ALM): step $\gamma \in (0, 2)$

$$\begin{cases} (x^{k+1}, y^{k+1}) \leftarrow \arg\min_{x,y} \{\mathcal{L}_{\mathcal{A}}(x, y, \lambda^k) : x \in \mathcal{X}, y \in \mathcal{Y}\} \\ \lambda^{k+1} \leftarrow \lambda^k - \gamma\beta\,(Ax^{k+1} + By^{k+1} - b) \end{cases}$$

Hestines-69, Powell-69, Rockafellar-73

(It does not explicitly use 2-separability)

RICE

Classic Alternating Direction Method (交替方向法)

Replace joint minimization by alternating minimization once:

$$\min_{x,y} \mathcal{L}_\mathcal{A} \;\approx\; (\min_x \mathcal{L}_\mathcal{A}) \oplus (\min_y \mathcal{L}_\mathcal{A})$$

(AL)ADM: step $\gamma \in (0, 1.618)$

$$\begin{cases} x^{k+1} \leftarrow \arg\min_x \{\mathcal{L}_\mathcal{A}(x, y^k, \lambda^k) : x \in \mathcal{X}\} \\ y^{k+1} \leftarrow \arg\min_y \{\mathcal{L}_\mathcal{A}(x^{k+1}, y, \lambda^k) : y \in \mathcal{Y}\} \\ \lambda^{k+1} \leftarrow \lambda^k - \gamma\beta\,(Ax^{k+1} + By^{k+1} - b) \end{cases}$$

It does use 2-separability: (“远交近攻”, “各个击破”)

- x-subproblem:

$$\min_x \; f_1(x) + \frac{\beta}{2}\|Ax - c_1(y^k)\|^2$$

- y-subproblem:

$$\min_y \; f_2(y) + \frac{\beta}{2}\|By - c_2(x^{k+1})\|^2$$

ADM overview (I)

ADM as we know today

- Glowinski-Marocco-75 and Gabay-Mercier-76
- Glowinski *at el.* 81-89, Gabay-83...

Connections before Aug. Lagrangian

- Douglas, Peaceman, Rachford (middle 1950's)
- operator splittings for PDE (a.k.a. ADI methods)

ADM overview (II)

After PDE, subsequent studies in optimization

- variational inequality, proximal-point, Bregman, ...
  (Eckstein-Bertsekas-92 ......)
- inexact ADM (He-Liao-Han-Yang-02 ......)
- Tseng-91, Fukushima-92, ...
- proximal-like, Bregman (Chen and Teboulle-93)
- ......

ADM had been used in optimization to some extent, but not as widely used to be called "main-stream" algorithm

RICE

ADM overview (III)

Recent Revival in Signal/Image/Data Processing

- ℓ_1-norm, total variation (TV) minimization
- convex, non-smooth, simple structures

Splitting + alternating:

- Wang-Yang-Yin-Z-2008, FTVd (TV code)
  (split + quadratic penalty, 2007)
  (split + quadratic penalty + multiplier in code, 2008)
- Goldstein-Osher-2008, split Bregman
  (split + quadratic penalty + Bregman, earlier in 2008)
- ADM ℓ_1-solver for 8 models: YALL1. Yang-Z-2010

Googled "split Bregman": "found 16,300 results".
Turns out that hot split Bregman = cool ALM

ADM Global Convergence

e.g., "*Augmented Lagrangian methods ...*" Fortin-Glowinski-83

Assumptions required by current theory:

- convexity over the entire domain
- separability for exactly 2 blocks, no more
- exact or high-accuracy minimization for each block

Strength:

- differentiability not required
- side-constraints allowed: $x \in \mathcal{X}, y \in \mathcal{Y}$

But

- why not 3 or more blocks?
- very rough minimization?
- rate of convergence?

Some Recent Applications

From PDE to:

Signal/Image Processing

Sparse Optimization

TV-minimization in Image Processing

TV/L^2 deconvolution model (Rudin-Osher-Fatemi-92):

$$\min_u \sum\nolimits_i \|D_i u\| + \frac{\mu}{2}\|Ku - f\|^2 \ \text{(sum all pixels)}$$

Splitting:

$$\min_{u,\mathbf{w}} \left\{ \sum \|\mathbf{w}_i\| + \frac{\mu}{2}\|Ku - f\|^2 : \mathbf{w}_i = D_i u, \forall i \right\}.$$

Augmented Lagrangian function $\mathcal{L}_{\mathcal{A}}(\mathbf{w}, u, \lambda)$:

$$\sum_i \left(\|\mathbf{w}_i\| - \lambda_i^\top (\mathbf{w}_i - D_i u) + \frac{\beta}{2}\|\mathbf{w}_i - D_i u\|^2 \right) + \frac{\mu}{2}\|Ku - f\|^2.$$

Closed formulas for minimizing w.r.t. $\mathbf{w}$ (shrinkage) and u (FFT)
(almost linear-time per iteration)

Shrinkage (or Soft Thresholding)

Solution to a simple optimization problem:

$$x(v, \mu) := \arg\min_{x \in \mathbb{R}^d} \|x\| + \frac{\mu}{2}\|x - v\|^2$$

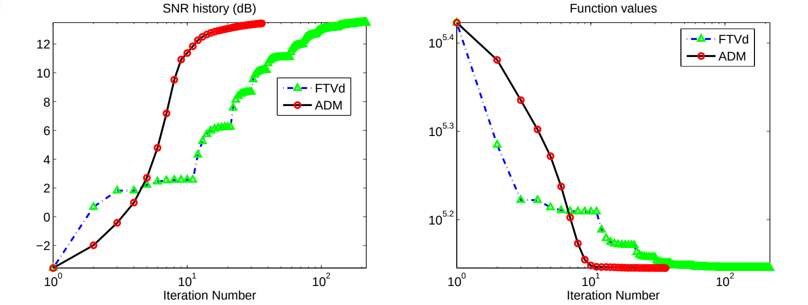
where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$, $v \neq 0$ and $\mu > 0$.

$$x(v, \mu) = \max\left(\|v\| - \frac{1}{\mu}, 0\right)\frac{v}{\|v\|}$$

This formula was used at least 30 years ago.

RICE

Multiplier helps: Penalty vs. ADM



Matlab package `FTVd` (Wang-Yang-Yin-Z, 07∼09):
`http://www.caam.rice.edu/~optimization/L1/ftvd/`
(v1-3 use Quadratic penalty, v4 applies ADM.

Orders of magnitude faster than PDE-based methods.

Key: "splitting-alternating" takes advantage of the structure.
Use of multiplier brings further speedup.

RICE

Example: Cross-channel blur + Gaussian noise

FTVd: $\min_u \mathrm{TV}(u) + \mu \|Ku - f\|_2^2$, sizes 512^2 and 256^2

Original

Blurry&Noisy. SNR: 8.01dB

FTVd: SNR: 19.54dB, t = 16.86s

Original

Blurry&Noisy. SNR: 6.70dB

FTVd: SNR: 18.49dB, t = 4.29s

(computation by Junfeng Yang)

ℓ_1-minimization in Compressive Sensing

Signal acquisition/compression: $A \in \mathbb{R}^{m \times n}$ $(m < n)$

$$b \approx Ax^* \in \mathbb{R}^m$$

where $x^* \in \mathbb{R}^n$ is sparse or compressible under a orthogonal transformation Ψ. ℓ_1 norm is used as the surrogate of sparsity.

8 signal recovery models: $A \in \mathbb{R}^{m \times n}$ $(m < n)$

1. $\min \|\Psi x\|_1, \text{ s.t. } Ax = b \quad (x \geq 0)$
2. $\min \|\Psi x\|_1, \text{ s.t. } \|Ax - b\|_2 \leq \delta \quad (x \geq 0)$
3. $\min \|\Psi x\|_1 + \mu \|Ax - b\|_2^2 \quad (x \geq 0)$
4. $\min \|\Psi x\|_1 + \mu \|Ax - b\|_1 \quad (x \geq 0)$

Can we solve these 8 model by $\leq$ 30 lines of 1 Matlab code?
YALL1 using ADM.

ℓ_1-minimization in Compressive Sensing (II)

Sparse signal recovery model: $A \in \mathbb{R}^{m\times n}$ $(m < n)$

$$\min \{\|x\|_1 : Ax = b\} \quad \overset{\text{dual}}{\Longleftrightarrow} \quad \max \left\{b^\top y : A^\top y \in [-1,1]^n\right\}$$

Add splitting z to "free" $A^\top y$ from the unit box:

$$\max \left\{b^\top y : A^\top y = z \in [-1,1]^n\right\}$$

ADM (1 of variants in Yang-Z-09): $AA^\top = I$ (common in CS)

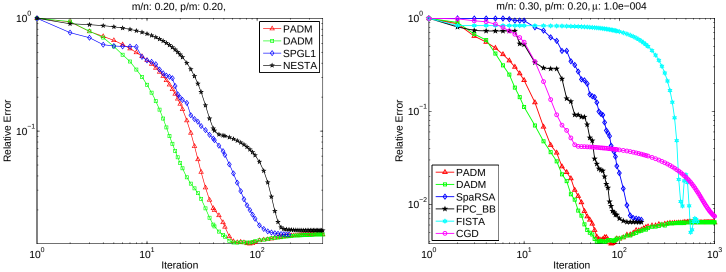
$$\begin{aligned} y &\leftarrow A(z - x) + b/\beta \\ z &\leftarrow \mathcal{P}_{[-1,1]^n}(A^\top y + x) \\ x &\leftarrow x - \gamma(z - A^\top y) \end{aligned}$$

RICE

Numerical Comparison

ADM solver package YALL1: `http://yall1.blogs.rice.edu/`

Compared codes: SPGL1, NESTA, SpaRSA, FPC, FISTA, CGD



(noisy measurements, average of 50 runs)

Nonasymptotically, ADMs showed the fastest speed of convergence in reducing error $\|x^k - x^*\|$.

Single Parameter β

In theory, $\beta > 0 \Longrightarrow$ convergence

How to choose the penalty parameter in practice?

In YALL1: **Make the subproblems scalar scale invariant**

- Scale A to "unit" size
- Scale b accordingly.
- $\beta = m/\|b\|_1$.

Optimal choice is still an open theoretical question.

Signal Reconstruction with Group Sparsity

Group-sparse signal $x = (x_1;\ \cdots;\ x_s), x_i \in \mathbb{R}^{n_i}, \sum_{i=1}^{s} n_i = n$

$$\min_x \sum_{i=1}^{s} \|x_i\|_2 \ \text{ s.t. } \ Ax = b.$$

Introduce splitting $y \in \mathbb{R}^n$,

$$\min_{x,y} \sum_{i=1}^{s} \|y_i\|_2 \ \text{ s.t. } \ y = x, \ Ax = b.$$

ADM (Deng-Yin-Z-10):

$$\begin{aligned} y &\leftarrow \text{shrink}(x + \lambda_1, 1/\beta) \quad \text{(group-wise)} \\ x &\leftarrow (I + A^T A)^{-1}((y - \lambda_1) + A^\top(b + \lambda_2)) \\ (\lambda_1, \lambda_2) &\leftarrow (\lambda_1, \lambda_2) - \gamma(y - x, Ax - b) \end{aligned}$$

Easy if $AA^T = I$; else take a steepest descent step in x (say).

Multi-Signal Reconstruction with Joint Sparsity

Recover a set of jointly sparse signals $X = [x_1 \ \cdots \ x_p] \in \mathbb{R}^{n \times p}$

$$\min_X \sum_{i=1}^n \|e_i^T X\| \ \text{ s.t. } \ A_j x_j = b_j, \ \forall j.$$

Assume $A_j = A$ for simplicity. Introduce splitting $Z \in \mathbb{R}^{p \times n}$,

$$\min_X \sum_{i=1}^n \|Z e_i\| \ \text{ s.t. } \ Z = X^T, \ AX = B.$$

ADM (Deng-Yin-Z-10):

$$\begin{array}{rcl} Z & \leftarrow & \text{shrink}(X^\top + \Lambda_1, 1/\beta) \ \ \text{(column-wise)} \\ X & \leftarrow & (I + A^T A)^{-1}((Z - \Lambda_1)^\top + A^\top (B + \Lambda_2)) \\ (\Lambda_1, \Lambda_2) & \leftarrow & (\Lambda_1, \Lambda_2) - \gamma(Z - X^T, AX - B) \end{array}$$

Easy if $AA^T = I$; else take a steepest descent step in X.

Extensions to Non-convex Territories

(as long as convexity exists in each direction)

Low-Rank/Sparse Matrix Models

Non-separable functions

More than 2 blocks

Matrix Fitting Models (I): Completion

Find low-rank Z to fit data $\{M_{ij} : (i,j) \in \Omega\}$

Nuclear-norm minimization is good, but SVDs are expensive.

Non-convex model (Wen-Yin-Z-09): find $X \in \mathbb{R}^{m\times k}$, $Y \in \mathbb{R}^{k\times n}$

$$\min_{X,Y,Z} \|XY - Z\|_F^2 \;\; \text{s.t.} \;\; \mathcal{P}_\Omega(Z - M) = 0$$

An SOR scheme:

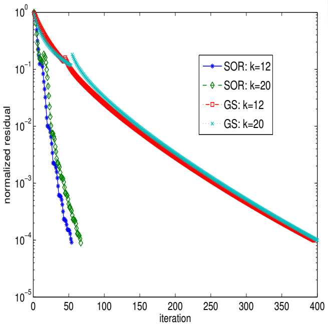
$$\begin{aligned} Z &\leftarrow \omega Z + (1-\omega)XY \\ X &\leftarrow \mathtt{qr}(ZY^\top) \\ Y &\leftarrow X^\top Z \\ Z &\leftarrow XY + \mathcal{P}_\Omega(M - XY) \end{aligned}$$

1 small QR ($m \times k$). No SVD. ω dynamically adjusted.

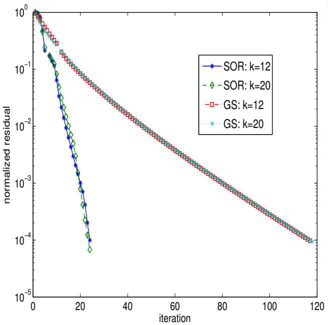
Much faster than nuclear-norm codes (when it is applicable)

RICE

Nonlinear GS vs SOR



(a) n=1000, r=10, SR = 0.08



(b) n=1000, r=10, SR=0.15

Alternating minimization, but no multiplier for storage reason

Is non-convexity a problem for global optimization of this problem?
— "Yes" in theory
— "Not really" in practice

RICE

Matrix Fitting Models (II): Separation

Given data $\{D_{ij} : (i,j) \in \Omega\}$,

Find low-rank Z so that difference $\mathcal{P}_\Omega(Z - D)$ is sparse

Non-convex Model (Shen-Wen-Z-10): $U \in \mathbb{R}^{m\times k}$, $V \in \mathbb{R}^{k\times n}$

$$\min_{U,V,Z} \|\mathcal{P}_\Omega(Z - D)\|_1 \ \text{ s.t. } \ Z - UV = 0$$

ADM scheme:

$$\begin{array}{rcl} U & \leftarrow & \mathtt{qr}((Z - \Lambda/\beta)V^\top) \\ V & \leftarrow & U^\top(Z - \Lambda/\beta) \\ \mathcal{P}_{\Omega^c}(Z) & \leftarrow & \mathcal{P}_{\Omega^c}(UV + \Lambda/\beta) \\ \mathcal{P}_\Omega(Z) & \leftarrow & \mathcal{P}_\Omega(\text{shrink}(\cdots) + D) \\ \Lambda & \leftarrow & \Lambda - \gamma\beta(Z - UV) \end{array}$$

— 1 small QR. No SVD. Faster.

— non-convex, 3 blocks. nonlinear constraint. convergence?

RICE

Nonnegative Matrix Factorization (Z-09)

Given $A \in \mathbb{R}^{n\times n}$, find $X, Y \in \mathbb{R}^{n\times k}$ $(k \ll n)$,

$$\min \|XY^\top - A\|_F^2 \ \text{ s.t. } \ X, Y \geq 0$$

Splitting:

$$\min \|XY^\top - A\|_F^2 \ \text{ s.t. } \ X = U_1, Y = U_2, U_1, U_2 \geq 0$$

ADM scheme:

$$\begin{aligned}
X &\leftarrow (AY + \beta(U_1 - \Lambda_1))(Y^\top Y + \beta I)^{-1} \\
Y^\top &\leftarrow (X^\top X + \beta I)^{-1}(X^\top A + \beta(U_2 - \Lambda_2)) \\
(U_1, U_2) &\leftarrow \mathcal{P}_+(X + \Lambda_1, Y + \Lambda_2) \\
(\Lambda_1, \Lambda_2) &\leftarrow (\Lambda_1, \Lambda_2) - \gamma(X - U_1, Y - U_2)
\end{aligned}$$

— cost/iter: 2 $(k \times k)$ linear systems plus matrix arithmetics

— better performance than Matlab built-in function "`nnmf`"

— non-convex, non-separable, 3 blocks: convergence?

RICE

Theoretical Convergence Results

A general setting

Local R-linear convergence

Global convergence for linear constraints

(Liu-Yang-Z, work in progress)

General Setting: Problem

Consider

$$\min_x f(x) \;\; \text{s.t.} \;\; c(x) = 0$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m (m < n)$ are $\mathcal{C}^2$-mappings.
Augmented Lagrangian:

$$\mathcal{L}_\alpha(x, y) \triangleq \alpha f(x) - y^T c(x) + \frac{1}{2}\|c(x)\|^2$$

Augmented saddle point system:

$$\begin{aligned} \nabla_x \mathcal{L}_\alpha(x, y) &= 0, \\ c(x) &= 0. \end{aligned}$$

Splitting and Iteration Scheme

$G : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a splitting of $F : \mathbb{R}^n \to \mathbb{R}^n$ if

$$G(x, x) \equiv F(x), \ \forall x \in \mathbb{R}^n.$$

e.g., if $A = L - R$, $G(x, x) \triangleq Lx - Rx \equiv Ax \triangleq F(x)$.

Let $G(x, x, y)$ be a splitting of $\nabla_x \mathcal{L}_\alpha(x, y)$ on x

Augmented saddle point system becomes

$$\begin{aligned} G(x, x, y) &= 0 \\ c(x) &= 0 \end{aligned}$$

A general Split (gSS) Scheme for Saddle-point Systems:

$$\begin{aligned} x^{k+1} &\leftarrow G(x, x^k, y^k) = 0 \\ y^{k+1} &\leftarrow y^k - \tau c(x^{k+1}) \end{aligned}$$

Block Jacobi for Square System $F(x) = 0$

Partition the system and variable into $s \leq n$ consistent blocks:

$$\boxed{F = (F_1, F_2, \cdots, F_s), \quad x = (x_1, x_2, \cdots, x_s)}$$

Block Jacobi iteration: given x^k, for $i = 1, 2, \ldots, s$

$$x_i^{k+1} \leftarrow F_i(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_s^k) = 0$$

$$\boxed{\text{or} \quad x^{k+1} \leftarrow G(x, x^k) = 0}$$

where

$$G(x, z) = \begin{pmatrix} F_1(x_1, z_2, \ldots, z_s) \\ \vdots \\ F_i(z_1, \ldots, x_i, z_{i+1}, \ldots, z_s) \\ \vdots \\ F_s(z_1, \ldots, x_s) \end{pmatrix}$$

RICE

Block Gauss-Seidel for Square System $F(x) = 0$

Block GS iteration: given x^k, for $i = 1, 2, \ldots, s$

$$x_i^{k+1} \;\leftarrow\; F_i(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_s^k) = 0$$

$$\text{or} \quad x^{k+1} \;\leftarrow\; G(x, x^k) = 0$$

where

$$G(x, z) = \begin{pmatrix} F_1(x_1, z_2, \ldots, z_s) \\ \vdots \\ F_i(x_1, \ldots, x_i, z_{i+1}, \ldots, z_s) \\ \vdots \\ F_s(x_1, \ldots, x_s) \end{pmatrix}$$

(SOR can be similarly defined.)

Splitting for Gradient Descent: $F(x) = \nabla f(x)$

Gradient descent method (with a constant step size):

$$x^{k+1} = x^k - \alpha F(x^k),$$

$$\text{or} \quad x^{k+1} \leftarrow G(x, x^k) = 0$$

where

$$G(x, z) = \frac{1}{\alpha} x - \left(\frac{1}{\alpha} z - F(z) \right).$$

— gradient descent iterations can be done block-wise
— block GS, SOR and gradient descent can be mixed
(e.g., 1st block: GS; 2nd block: gradient descent)

RICE

Assumptions

Let $\partial_i G(x, x, y)$ be the partial Jacobian of the splitting G w.r.t. the i-th argument, and $\partial_i G^* \triangleq \partial_i G(x^*, x^*, y^*)$ where x^* is a minimizer and y^* the associated multiplier.

Assumption 1. (2nd-order Sufficiency)
$f, c \in \mathcal{C}^2$, and $\alpha > 0$ is chosen so that

$$\nabla_x^2 \mathcal{L}_\alpha(x^*, y^*) \succ 0$$

Assumption 2. (Requirement on splitting)
$\partial_1 G$ is nonsingular in a neighborhood of (x^*, x^*, y^*), and

$$\rho\left([\partial_1 G^*]^{-1} \partial_2 G^*\right) < 1$$

(e.g., for gradient descent: $[\partial_1 G^*]^{-1} \partial_2 G^* = I - \alpha \nabla^2 f(x^*)$)

Assumptions are Reasonable

A1. 2nd-order sufficiency guarantees that $\alpha > 0$ exists so that

$$\alpha \left[\nabla^2 f(x^*) - \sum_i \hat{y}_i^* \nabla^2 c_i(x^*)\right] + A(x^*)^\top A(x^*) \succ 0$$

where $A(x) = \partial c(x)$. Note

$$\nabla_x \mathcal{L}_\alpha(x, y) = G(x, x, y) \implies \nabla_x^2 \mathcal{L}_\alpha^* = \partial_1 G^* + \partial_2 G^* \succ 0$$

A2. Any convergent linear splitting for matrices $\succ 0$ leads to a corresponding nonlinear splitting G satisfying

$$\rho\left([\partial_1 G^*]^{-1} \partial_2 G^*\right) < 1$$

Hence, **A2** is satisfied by block GS (i.e., ADM), SOR, gradient descent (with appropriate α) and their mixtures.

The Error System

Recall gSS:

$$\begin{array}{rcl} x^{k+1} & \leftarrow & G(x, x^k, y^k) = 0 \\ y^{k+1} & \leftarrow & y^k - \tau c(x^{k+1}) \end{array}$$

Using Implicit Function Theorem, we derive an error system

$$e^{k+1} = M^*(\tau) e^k + o(\|e^k\|)$$

where $e^k \triangleq (x^k, y^k) - (x^*, y^*)$,

$$M^*(\tau) = \begin{bmatrix} -[\partial_1 G^*]^{-1} \partial_2 G^* & [\partial_1 G^*]^{-1} A^{*\top} \\ \tau A^* [\partial_1 G^*]^{-1} \partial_2 G^* & I - \tau A^* [\partial_1 G^*]^{-1} A^{*\top} \end{bmatrix}$$

Key Lemma. (Z-2010) Under Assumptions 1-2, there exists $\eta > 0$ such that $\rho(M^*(\tau)) < 1$ for all $\tau \in (0, 2\eta)$.

Convergence: $\tau \in (0, 2\eta)$

Theorem [Local convergence].
There exists an open neighborhood U of a KKT point (x^*, y^*) such that for any $(x^0, y^0) \in U$, the sequence $\{(x^k, y^k)\}$ generated by gSS stays in U and converges to (x^*, y^*).

Theorem [R-linear rate].
The asymptotic convergence rate of gSS is R-linear with R-factor $\rho(M^*(\tau))$, i.e.,

$$\limsup_{k\to\infty} \|(x^k, y^k) - (x^*, y^*)\|^{1/k} = \rho(M^*(\tau)).$$

— These follow from the Key Lemma and Ortega-Rockoff-70.

Corollary [quadratic case].
If f is quadratic and c is affine, then $U = \mathbb{R}^n \times \mathbb{R}^m$ and the convergence is globally Q-linear with Q-factor $\rho(M^*(\tau))$.

Global Convergence: Linear Constraints

$$\min_x f(x_1, \cdots, x_p), \text{ s.t. } \sum A_i x_i = b$$

1st-order optimality or saddle point system:

$$\begin{aligned} \nabla f(x) &= A^\top y \\ Ax - b &= 0 \end{aligned}$$

Augmented saddle point system:

$$\begin{aligned} \nabla f(x) + \beta A^\top (Ax - b) &= A^\top y \\ y - \tau\beta(Ax - b) &= y \end{aligned}$$

Splittings ($F(x) = G(x, x)$) can be applied to the 1st equation.

- Block Jacobi type give block diagonal split
- ADM: a block Gauss-Seidel type split

RICE

Global Convergence (preliminary)

$$\min_x f(x_1, \cdots, x_p), \text{ s.t. } \sum A_i x_i = b$$

f is separable if $f(x_1, \cdots, x_p) = \sum_i^p f_i(x_i)$. In this case, the Hessian is block diagonal.

Block Jacobi scheme:
If $f \in \mathcal{C}^2$ is separable, and each

$$\nabla^2 f_i(x_i) + \beta A_i^T A_i \succeq \epsilon I,$$

$\nabla_x^2 \mathcal{L}_\alpha$ is uniformly block diagonally dominant, then the block Jacobi scheme converges to a KKT point.

It can be extended to more general settings (GS, ...) under further assumptions (still under scrutiny).

The number of blocks can be arbitrary without modification
Other multi-block extensions exist with convexity and algorithm modifications (He and Yuan *et al*).

Summary: ADM $\simeq$ Splitting + Alternating

A simple yet effective approach to exploiting structures:

- bypasses non-differentiability
- enables very cheap iterations
- has at least an R-linear rate
- great versatility, good efficiency

Many issues remain. Convergence theory needs more work.

References on Codes

- (FISTA) A. Beck, and M. Teboulle, "A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems", *SIAM J. Imag. Sci.*, 2:183-202, 2009.
- (NESTA) J. Bobin, S. Becker, and E. Candes, "NESTA: A Fast and Accurate First-order Method for Sparse Recovery", *TR, CalTech*, April 2009.
- (SPGL1) M. P. Friedlander and E. van den Berg, "Probing the Pareto frontier for basis pursuit solutions", *SIAM J. Sci. Comput.*, 31(2):890–912, 2008.
- (FPC) E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for l1-minimization: Methodology and convergence", *SIAM J. Optim*, 19(3):1107–1130, 2008.
- (IALM) Z. Lin, M. Chen, L. Wu, and Y. Ma, "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices", *TR UIUC, UILU-ENG-09-2215*, Nov. 2009.
- (FPCA) S. Ma, D. Goldfarb and L. Chen, "Fixed Point and Bregman Iterative Methods for Matrix Rank Minimization", *Math. Prog.*, to appear.
- (APGL) K.-C. Toh, and S. W. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems", *Pacific J. Optimization*.
- (SpaRSA) S. Wright, R. Nowak, M. Figueiredo, "Sparse reconstruction by separable approximation", *IEEE Trans Signal Process.*, 57(7):2479–2493, 2009.
- (CGD) S. Yun, and K.-C. Toh, "A coordinate gradient descent method for L1-regularized convex minimization", *Computational Optimization and Applications*, to appear.

RICE