Recent Advances in Alternating Direction Methods: Practice and Theory

Yin Zhang

Department of Computational and Applied Mathematics
Rice University, Houston, Texas, U.S.A.

"Numerical Methods for Continuous Optimization"
IPAM Workshop at UCLA
Los Angeles, October 11, 2010

Outline:

- Alternating Direction Method (ADM)
  — algorithm, history, convergence
- Some Recent Applications and Extensions
  — convex and non-convex models
- Local Convergence and Rate
  — general equality constrained problem
  — general splitting and stationary iterations
- Summary

The talk involves contributions from (in random order):

Junfeng Yang, Zaiwen Wen, Yilun Wang, Chengbo Li, Yuan Shen, Wotao Yin

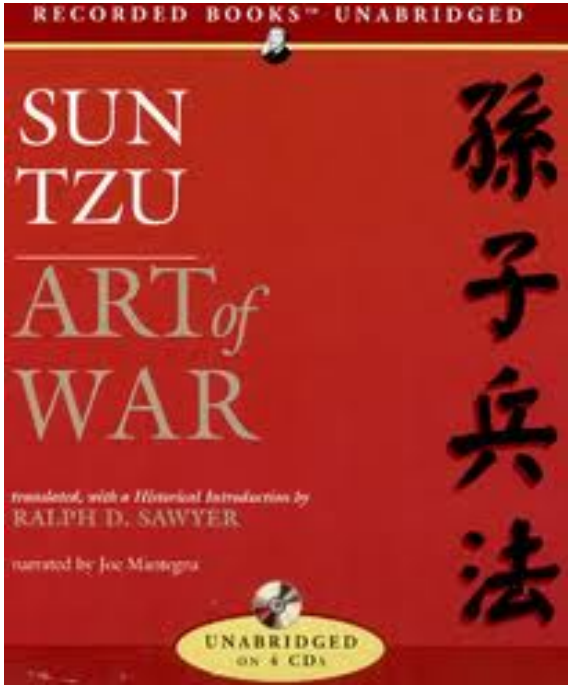
RICE

Introduction to ADM

Idea:

"Divide and Conquer." — Julius Caesar (100-44 BC)

"远交近攻", "各个击破". — Sun-Tzu (400 BC)



Ms. Paris Hilton reading Sun Tzu's Art of War book translated by Dr. Ralph Sawyer. Brava!

Classic Alternating Direction Method

Convex program with structure (2-block separability):

$$\min_{x,y} \{f_1(x) + f_2(y) : Ax + By = b, x \in \mathcal{X}, y \in \mathcal{Y}\}.$$

Augmented Lagrangian (AL):

$$\mathcal{L}_{\mathcal{A}}(x, y, \lambda) = f_1(x) + f_2(y) + \frac{\beta}{2}\left(\|Ax + By - b - \lambda\|^2 - \|\lambda\|^2\right)$$

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

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

"An inexact version" of AL Method. Hestines-69, Powell-69

(AL Iterations = Bregman Iterations by Yin-Osher-Goldfarb-Darbon-08)

ADM overview

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)

Some subsequent studies

- variational inequality, proximal point (Eckstein-Bertsekas-92)
- inexact ADM (He-Liao-Han-Yang-02), ...

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

- some assumptions are restrictive
- no rate of convergence result

Some Recent Applications

From PDE to:

Signal/Image Processing

Sparse Optimization

ℓ_1-minimization in Compressive Sensing

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

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

Split A from the box:

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

ADM (1 of variants in Yang-Z-09):

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

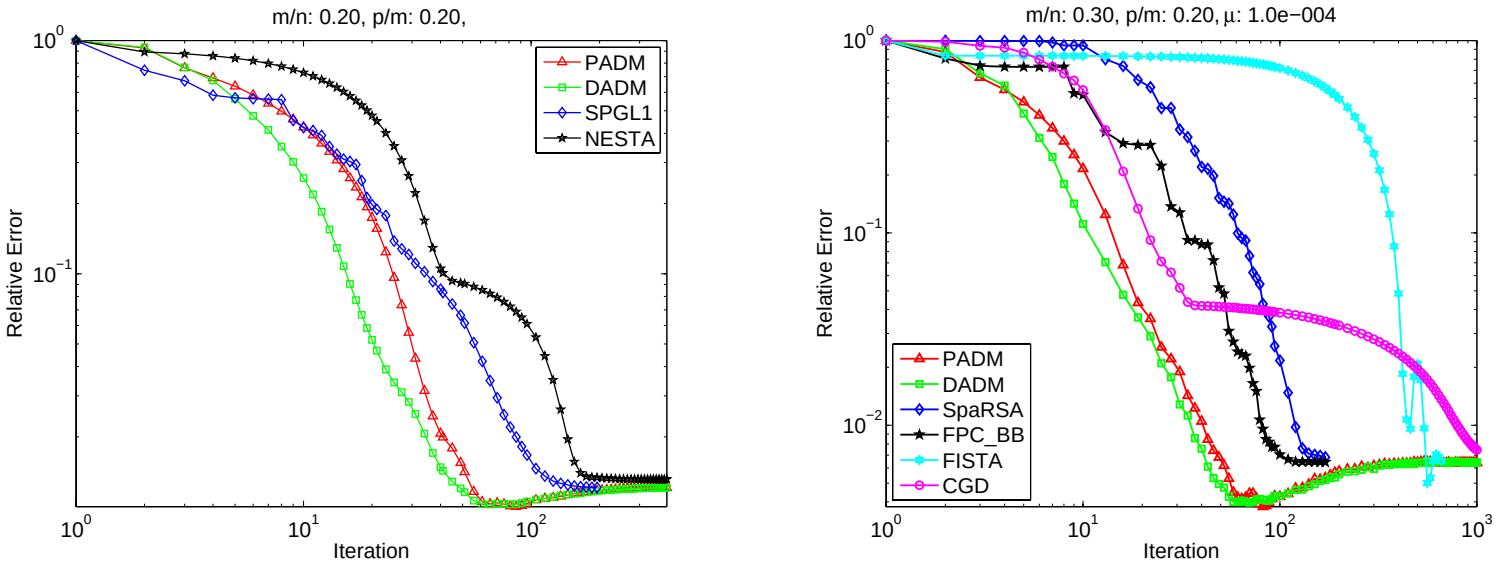
— Most efficient when $AA^\top = I$ (common in CS).

— Use a steepest descent step for y if $AA^\top \neq I$
(still good in CS because of well-conditioned A)

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 $\|x^k - x^*\|$.

TV-minimization in Image Processing

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

$$\min_u \sum\nolimits_i \|D_i u\| + \frac{\mu}{2}\|Ku - f\|^2,$$

which is equivalent to:

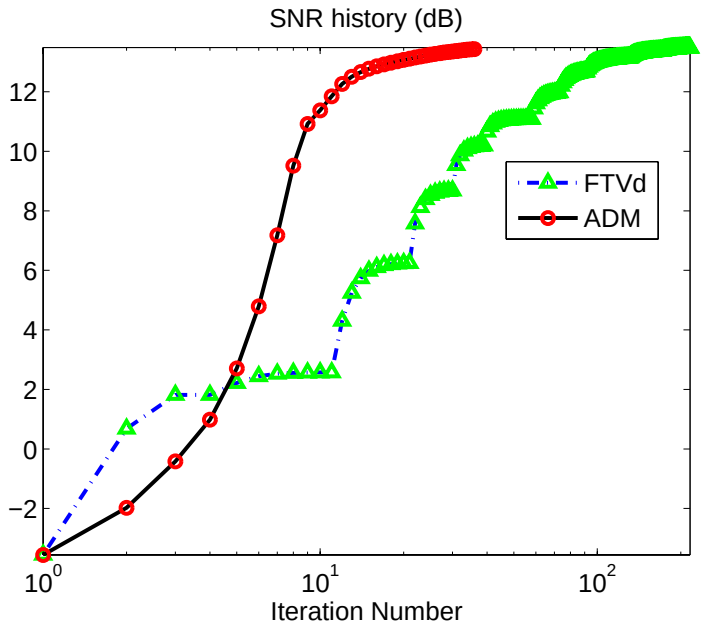
$$\min_{u,\mathbf{w}} \left\{ \sum\nolimits_i \|\mathbf{w}_i\| + \frac{\mu}{2}\|Ku - f\|^2 : \mathbf{w}_i = D_i u, \forall i \right\}.$$

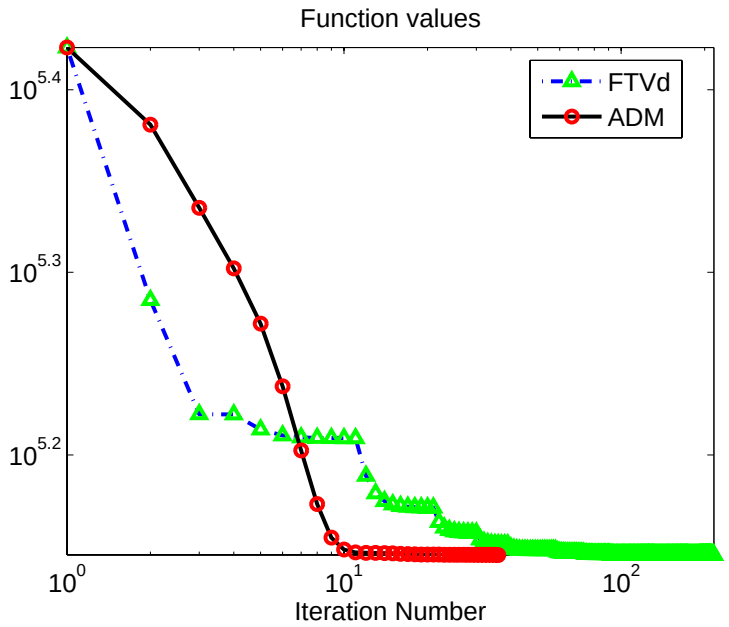
Augmented Lagrangian function $\mathcal{L}_{\mathcal{A}}(\mathbf{w}, u, \lambda)$:

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

Exact minimization for $\mathbf{w}$ (shrinkage) and u (FFT)
(e.g., Yang-Yin-Z-Wang-08 ($\lambda_i = 0$), Goldstein-Osher-08, Esser-09).

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)

Multi-Signal Reconstruction with Joint Sparsity

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

$$\min_X \sum_{i=1}^{n} \|e_i^T X\|_2 \ \text{ s.t. } \ A_j x_j = b_j, \ \forall j.$$

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

$$\min_X \sum_{i=1}^{n} \|Z e_i\|_2 \ \text{ s.t. } \ Z = X^T, \ AX = B.$$

ADM;

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

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

RICE

A Sample of Recent Works on Convex Programs

- D. Goldfarb, S. Ma, and K. Scheinberg, "Fast Alternating Linearization Methods for Minimizing the Sum of Two Convex Functions", TR *Columbia*, 2010.
- M. V. Afonso, J. Bioucas-Dias, and M. Figueiredo, "A fast algorithm for the constrained formulation of compressive image reconstruction and other linear inverse problems", Arxiv, 2009.
- S. Setzer, "Split Bregman Algorithm, Douglas-Rachford Splitting and Frame Shrinkage", Proc. 2nd International Conference on Scale Lecture Notes in Computer Science, 2009.

RICE

Extensions to Non-convex Territories

Low-Rank/Sparse Matrix Models

Non-convex, 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 $X \in \mathbb{R}^{m\times k}$, $Y \in \mathbb{R}^{k\times n}$

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

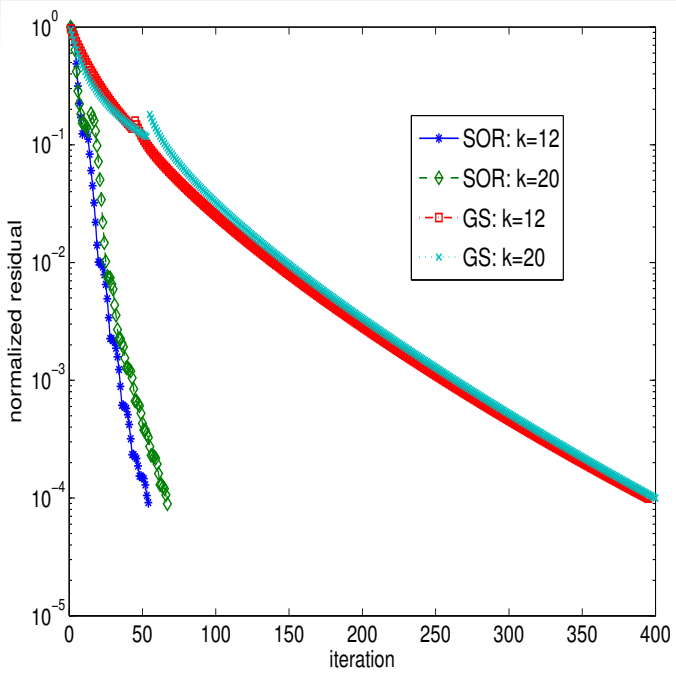
An SOR scheme:

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

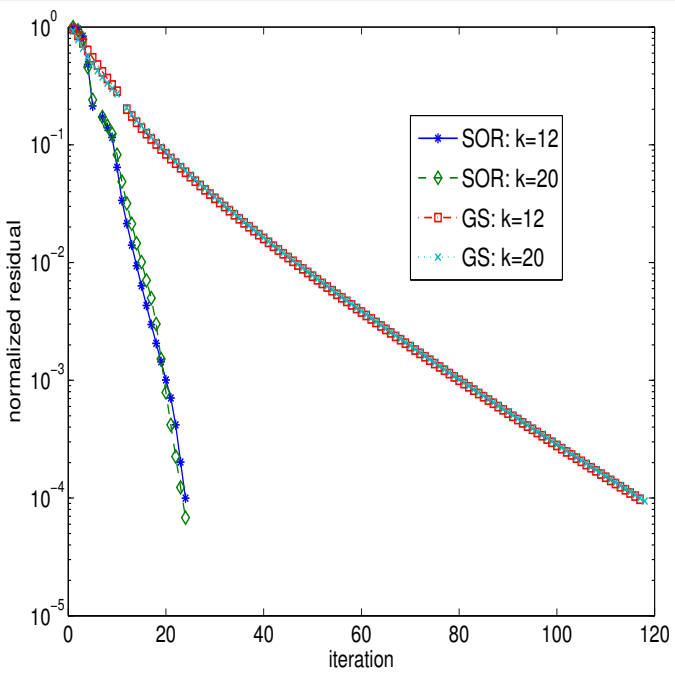
1 small QR ($m \times k$). No SVD.

Much faster than SVD-based codes (FPCA, APGL, ...)

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

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:

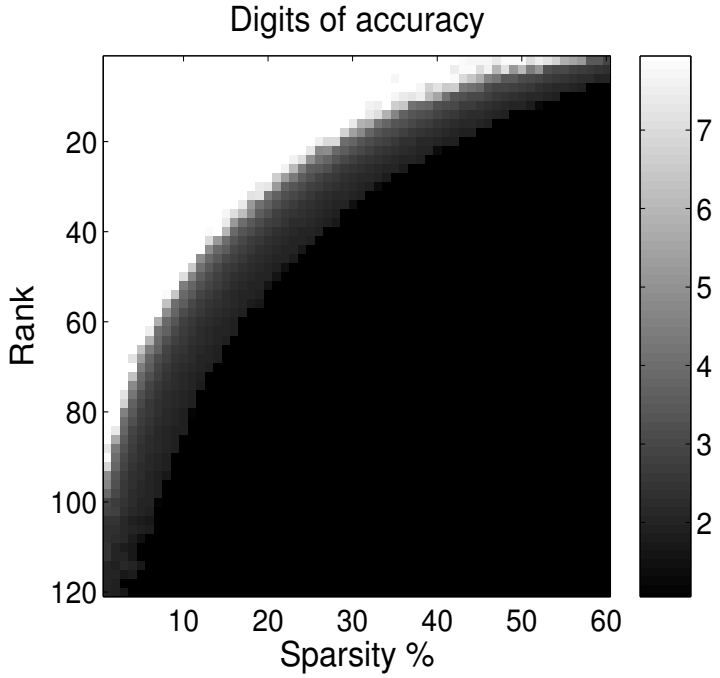
$$\begin{array}{rcl} U & \leftarrow & \mathtt{qr}((Z - \Lambda/\beta)V^\top) \\ V & \leftarrow & U^\top(Z - \Lambda/\beta) \\ \mathcal{P}_{\Omega^c}(Z) & \leftarrow & \mathcal{P}_{\Omega^c}(UV + \Lambda/\beta) \\ \mathcal{P}_\Omega(Z) & \leftarrow & \mathcal{P}_\Omega(\text{shrink}(\cdots) + D) \\ \Lambda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

— **1 small QR. No SVD. Faster.**

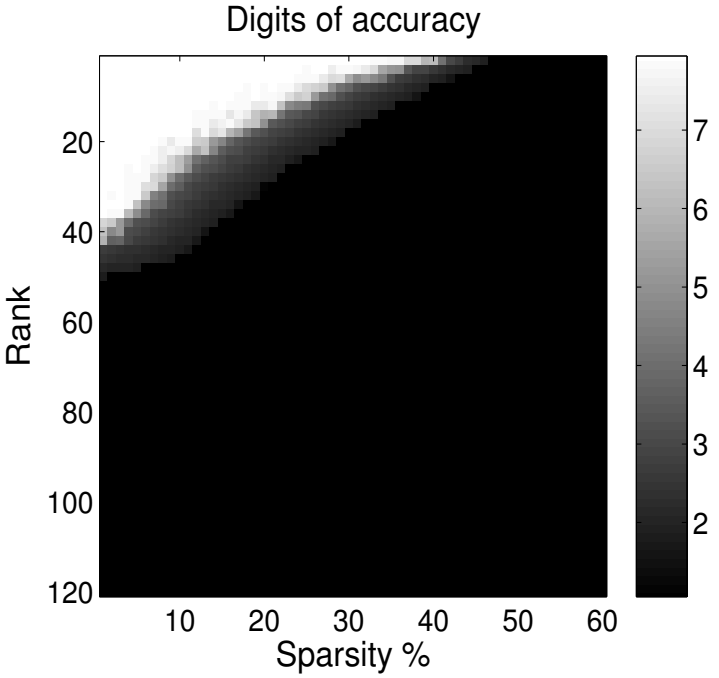
— non-convex, 3 blocks. nonlinear constraint. convergence?

RICE

Matrix Separation: recoverability can be better



(a) LMaFit (our solver)



(b) IALM (nuclear-norm)

In $S + Z = D$, when S dominates Z in magnitude, our model has worse recoverability. So the 2 models complement each other.

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:

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

— 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

Minimization on Stiefel Manifold

$$\min_{X \in \mathcal{X}} f(X) \ \text{ s.t. } \ X^\top X = I$$

Splitting:

$$\min_{X \in \mathcal{X}} f(X) \text{ s.t. } X - Z = 0, \ Z^T Z = I$$

ADM framework:

$$\begin{aligned} X &\leftarrow \text{argmin } \{f(X) + \frac{\beta}{2}\|X - (Z + \Lambda)\|_F^2 : X \in \mathcal{X}\} \\ Z &\leftarrow \text{argmin } \{\|Z - (X - \Lambda)\|_F^2 : Z^T Z = I\} \\ \Lambda &\leftarrow \Lambda - \gamma(X - Z) \end{aligned}$$

— X-subproblem: easier in many cases
— Z-subproblem: close-form solution with an SVD
— a more complex splitting necessary to avoid SVD

Nonlinear Data-Fitting with Regularization

$$\min_v R(Dv) + \frac{\mu}{2}\|K(v) - d\|^2.$$

Introduce splitting u,

$$\min_{u,v} R(u) + \frac{\mu}{2}\|K(v) - d\|^2 \ \text{ s.t. } \ u = Dv.$$

A general ADM scheme:

$$\begin{aligned} u &\leftarrow \operatorname*{argmin}_u \ R(u) + \frac{\beta}{2}\|u - (Dv + \lambda)\|^2, \\ v &\leftarrow \operatorname*{argmin}_v \ \mu\|K(v) - d\|^2 + \beta\|Dv - (u - \lambda)\|^2, \\ \lambda &\leftarrow \lambda - \gamma(u - Dv), \end{aligned}$$

— u-subproblem: easy for many $R(\cdot)$ (TV, ℓ_1, ...)
— v-subproblem: regularized by a Tikhonov term
— v-subproblem: Gauss-Newton step could suffice

RICE

Some Theoretical Results

A general setting

Local R-linear convergence

(Joint work in progress with Junfeng Yang)

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 generalized ADM (gADM) scheme:

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

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

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

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

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)

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

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

but $|\lambda| = 1$ is allowed only for $\lambda = 1$.
(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 gADM:

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

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 $\tau \in (0, 2\eta)$.

RICE

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

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

Theorem [R-linear rate].
The asymptotic convergence rate of gADM 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 [Linear 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 $\rho(M^*(\tau))$.

Summary: ADM $\simeq$ Splitting + Alternating

- powerful tool for constructing inexpensive iterations
- great versatility, computationally proven effectiveness
- convergence theory extended, many issues remain

Table: assessment of the obtained results

strength	**weakness**
non-convex functions	local convergence only
non-separable functions	smooth functions only
any number of blocks	no side-constraints
mixed strategies covered	—
rate of convergence	—

Our results extend the classic ADM in several aspects, but do not recover the existing results except for quadratic programs.

RICE

Open Question:

Why or when is non-convexity NOT a problem?

Thanks!

Acknowledgments: NSF, ONR, IPAM

References

- (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