Beyond Heuristics: Applying Alternating Direction Method of Multipliers in Nonconvex Territory

Xin Liu(刘歆)

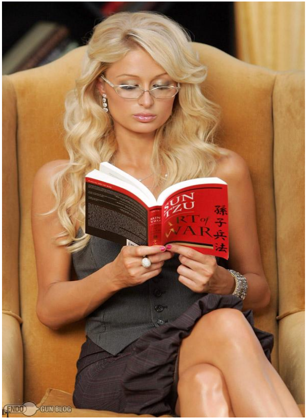
State Key Laboratory of Scientific and Engineering Computing
Institute of Computational Mathematics and Scientific/Engineering Computing
Academy of Mathematics and Systems Science
Chinese Academy of Sciences, China

2012 International Workshop on Signal Processing, Optimization, and Control
USTC, Hefei, Anhui, China

Outline

Section I. Introduction and Application

An Ancient Strategy



- “远交近攻，各个击破”，
  “分而治之”
  ——《孙子兵法》(《SUN TZU, ART OF WAR》)
  孙子(535 – 470 BC)
- “Divide et impera”
  Julius Caesar (100 – 44 BC)

Mathematical Point of View: Split and Alternate

Splitting Techniques

Case 1: Nondifferentiable Term

$$\min f(x) + g(Bx)$$

$$\Downarrow$$

$$\min f(x) + g(y) \qquad \text{s.t.} \;\; Bx - y = 0.$$

Case 2: Highly Nonconvex

$$\min f(g(x))$$

$$\Downarrow$$

$$\min f(y) \qquad \text{s.t.} \;\; g(x) - y = 0.$$

Case 3: Inconsistent Objective and Constraint

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

$$\Downarrow$$

$$\min f(x) \qquad \text{s.t.} \;\; c(y) = 0, \; x = y.$$

Splitting Instances

Instance 1: Compressive Sensing

$$\min \|Wx\|_1 + \frac{\mu}{2}\|Ax - b\|_2^2$$

$$\Downarrow$$

$$\min \|y\|_1 + \frac{\mu}{2}\|Ax - b\|_2^2 \quad \text{s.t.} \quad Wx - y = 0.$$

Instance 2: Nonlinear ℓ_1 Minimization

$$\min \ \|f(x)\|_1.$$

$$\Downarrow$$

$$\min \ \|y\|_1 \quad \text{s.t.} \ f(x) = y.$$

Instance 3: Dual Problem of Compressive Sensing (Yang-Zhang 2009)

$$\min \; -b^{\mathsf{T}}y + \frac{1}{2\mu}\|y\|_2^2 \qquad \text{s.t. } \; \|W^{-\mathsf{T}}A^{\mathsf{T}}y\|_\infty \le 1.$$

$$\Downarrow$$

$$\min \; -b^{\mathsf{T}}y + \frac{1}{2\mu}\|y\|_2^2 \qquad \text{s.t. } \; \|z\|_\infty \le 1, \quad z = W^{-\mathsf{T}}A^{\mathsf{T}}y.$$

Equality Constrained Problems

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

Augmented Lagrangian Function (Henstenes 1969, Powell 1969, Rockafellar 1973)

$$\mathcal{L}_\beta(x, \lambda) = f(x) - \lambda^\top c(x) + \frac{\beta}{2}\|c(x)\|_2^2.$$

Augmented Lagrangian Method

$$\text{ALM} : \begin{cases} x^{k+1} \leftarrow \arg\min \mathcal{L}_\beta(x, \lambda^k); \\ \lambda^{k+1} \leftarrow \lambda^k - \tau \beta\, c(x^{k+1}); \\ \text{update } \beta \text{ if necessary }. \end{cases}$$

Augmented Lagrangian Method (Cont'd)

Problems with Equality Constraints

$$\min_{x\in\Omega} f(x) \qquad \text{s.t.} \;\; c(x) = 0.$$

Augmented Lagrangian Method – Extension

$$\text{ALM}: \begin{cases} x^{k+1} \leftarrow \arg\min_{x\in\Omega} \mathcal{L}_\beta(x, \lambda^k); \\ \lambda^{k+1} \leftarrow \lambda^k - \tau\beta\, c(x^{k+1}); \\ \text{update } \beta \text{ if necessary }. \end{cases}$$

Block Structure

$$\{x \mid x \in \Omega\} = \bigcap_{i=1}^{p} \{x \mid x_i \in \Omega_i\}.$$

(Augmented Lagrangian) Alternating Direction Method (of Multiplier)

(Glowinski-Marocco 1975, Gabay-Mercier 1976, ...)

$$\text{ADMM}: \begin{cases} x_1^{k+1} \leftarrow \arg\min_{x_1 \in \Omega_1} \mathcal{L}_\beta(x_1, x_2^k, \dots, x_p^k, \lambda^k); \\ x_2^{k+1} \leftarrow \arg\min_{x_2 \in \Omega_2} \mathcal{L}_\beta(x_1^{k+1}, x_2, x_3^k, \dots, x_p^k, \lambda^k); \\ \dots\dots \\ x_p^{k+1} \leftarrow \arg\min_{x_p \in \Omega_p} \mathcal{L}_\beta(x_1^{k+1}, \dots, x_{p-1}^{k+1}, x_p, \lambda^k); \\ \lambda^{k+1} \leftarrow \lambda^k - \tau\beta\, c(x_1^{k+1}, \dots, x_p^{k+1}). \end{cases}$$

Phase Retrieval (Wen-Yang-L.)

- X-ray crystallography, transmission electron microscopy
- **Original model:**

$$\min_{\hat{\psi} \in \mathbb{C}^n} \sum_{i=1}^{k} \frac{1}{2} \left\| |\mathcal{F} Q_i \hat{\psi}| - b_i \right\|_2^2 .$$

- **Reformulation:**

$$\min_{\hat{\psi} \in \mathbb{C}^n, z \in \mathbb{C}^{m \times k}} \sum_{i=1}^{k} \frac{1}{2} \| |z_i| - b_i \|_2^2 \quad \text{s.t.} \ z_i = \mathcal{F} Q_i \hat{\psi}, \quad i = 1, ..., k.$$

- **Augmented Lagrange function:**

$$\mathcal{L}_\beta(z_i, \psi, y_i) = \sum_{i=1}^{k} \left(\frac{1}{2} \| |z_i| - b_i \|_2^2 + y_i^* (\mathcal{F} Q_i \psi - z_i) + \frac{\beta}{2} \| \mathcal{F} Q_i \psi - z_i \|_2^2 \right).$$

Portfolio Optimization (Wen-Peng-L.-Bai-Sun)

- Asset Allocation under the Basel Accord Risk Measures (Value-at-Risk) – integer programming
- **Original model:**

$$\min_{u \in \mathcal{U}_{r_0}} (-\tilde{R}u)_{(p)},$$

where $\mathcal{U}_{r_0} = \{u \in \mathbb{R}^d \mid \mu^\top u \geq r_0, \mathbf{1}^\top u = 1, u \geq 0\}$; $(\cdot)_{(p)}$ refers to the p-th smallest component of a vector.

- **Reformulation:**

$$\min_{u \in \mathcal{U}_{r_0},\, x \in \mathbb{R}^n} x_{(p)} \qquad \text{s.t.} \quad x + \tilde{R}u = 0.$$

- **Augmented Lagrange function:**

$$\mathcal{L}_\beta(x, u, \lambda) := x_{(p)} - \lambda^\top (x + \tilde{R}u) + \frac{\beta}{2}\|x + \tilde{R}u\|^2.$$

Applications III

Matrix Factorization (Zhang et al.)

- Nonnegative matrix factorization, structure enforcing matrix factorization
- **Original model:**

$$\min_{W\in\mathbb{R}^{m\times k},\ H\in\mathbb{R}^{n\times k}} \|A - WH^{\mathsf{T}}\|_{\mathsf{F}}^2 \quad \text{s.t.}\ \ W \in \mathbb{T}_1,\ H \in \mathbb{T}_2,$$

  where $\mathbb{T}_1$, $\mathbb{T}_2$ can be $\{X \mid X^{\mathsf{T}}X = I\}$, or $\{X \mid X \geq 0\}$, or any other matrix sets allowing 'easy projection'

- **Reformulation:**

$$\min_{W,\ H,\ S_1\in\mathbb{T}_1,\ S_2\in\mathbb{T}_2} \|A - WH^{\mathsf{T}}\|_{\mathsf{F}}^2 \quad \text{s.t.}\ \ W = S_1,\ H = S_2.$$

- **Augmented Lagrange function:**

$$\begin{aligned}\mathcal{L}_{(\beta_1,\,\beta_2)}(W, H, S_1, S_2, \Lambda) \;&=\; \|A - WH^{\mathsf{T}}\|_{\mathsf{F}}^2 - \Lambda_1 \bullet (W - S_1) \\ -\Lambda_2 \bullet (H - S_2) \;&+\; \frac{\beta_1}{2}\cdot\|W - S_1\|_{\mathsf{F}}^2 + \frac{\beta_2}{2}\cdot\|H - S_2\|_{\mathsf{F}}^2.\end{aligned}$$

Section II. Theoretical Results

Brief Introduction

Intuition

- "Splitting" brings easy subproblem
- "Splitting" induces equality constraint – Augmented Lagrange
- "Alternating" solves the split targets in turn
- From line search to ADM
  - Line search based optimization - one dimensional subspace
  - Subspace method - multi-dimensional subspace
  - ADMM - high-order subspaces

Convergence Based on Strict Conditions

- Two blocks, joint convexity, separability (Gabay-Mercier 1976)
- Multiple blocks, variant versions (He, Yuan et al., Goldfarb and Ma, etc.)
  - complexity
  - acceleration
  - customization
- Two blocks, linear convergence rate (Yin-Deng 2012)

Towards a General Scheme

– nonconvex and nonseparable cases

- Local convergence and rate (Yang-L.-Zhang)
- Global convergence (L.-Yang-Zhang)
  - under some assumptions (ongoing)
  - special case: multiple blocks,
    separable + strongly convex + second order differentiable

Nonlinear Splitting and Iteration Scheme

- Original Nonlinear System: $F : \mathbb{R}^n \to \mathbb{R}^n$
- Splitting: $G : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$
  i.e. $G(x, x) := L(x) - R(x) \equiv F(x)$. $\partial_1 G \triangleq \partial_x G$, and $\partial_2 G \triangleq -\partial_x G$.
- Consider $G(x, x, \lambda)$ to be a splitting of $F := \nabla_x \mathcal{L}_\beta(x, \lambda)$
- A generalized ADMM scheme:

$$\text{GADMM} : \begin{cases} x^{k+1} \leftarrow G(x, x^k, \lambda^k) = 0; \\ \lambda^{k+1} \leftarrow \lambda^k - \tau\beta\, c(x^{k+1}). \end{cases}$$

Local Convergence Result

Error System

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

where

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

Local convergence:

- $e^k \triangleq ((x^k - x^*)^\top, (\lambda^k - \lambda^*)^\top)^\top$
- Implicit Function Theorem + Taylor Expansion
- Assumptions: $\nabla_{xx}\mathcal{L}_\beta(x^*, \lambda^*) \succ 0$ and $\nabla c(x^*)$ full row rank
- Results:
  - local convergence: $\exists\, \eta > 0, \quad \rho(M(\tau)) < 1, \quad \forall\, \tau \in (0, \eta)$;
  - R-linear rate: $\rho(M(\tau))$.

Global Convergence

Relative Error System

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

where

$$M(\tau)^k = \begin{bmatrix} [\bar{\partial}_1 G_L^k]^{-1} \bar{\partial}_2 G_U^k & [\bar{\partial}_1 G_L^k]^{-1} A^\top \\ -\tau A [\bar{\partial}_1 G_L^k]^{-1} \bar{\partial}_2 G_U^k & I - \tau A [\bar{\partial}_1 G_L^k]^{-1} A^\top \end{bmatrix}$$

Global convergence:

- $e^k \triangleq ((x^k - x^{k-1})^\top, (\lambda^k - \lambda^{k-1})^\top)^\top$
- Mean Value Theorem + Average Hessian $(\bar{\partial}_1 G_L^k, \bar{\partial}_2 G_U^k)$
- Difficulty: **non-stationary iteration**

  $\mathcal{L}_\beta$ strongly convex and $\nabla \mathcal{L}_\beta$ is Lipschitz continuous $\Rightarrow$

  $\rho(M(\tau)^k) \leq 1 - \epsilon$ $(\forall k)$ $\nLeftrightarrow$ global convergence

Global Convergence (Cont'd)

ℓ_2 Restriction (ongoing)

- $\|M(\tau)^k\|_2 \leq 1 - \epsilon$ $(\forall k)$
- Assumptions:
  - linear constraints • block-wise convexity
  - second order differentiability • block diagonal dominance
- Result: global convergence

Special Case

- Assumptions:
  - separability: $\bar{\partial}_2 G_U^k$ constant, $\bar{\partial}_1 G_L^k$ non-stationary in block diagonal
  - strongly convexity • linear constraints • second order differentiability
- Result:
  - $\exists \bar{\beta} > 0$ and $\exists \eta > 0$;
  - global convergence, $\forall \beta \in (0, \bar{\beta})$, $\forall \tau \in (0, \eta)$.

Section III. Conclusion and Future works

Conclusion

- Powerful tool for hard optimization problem with structure;
- Lack of convergence results for nonconvex problems;
- Excellent performance in practice.

Future Works

- There is still room for further improvement of the algorithm;
- Convergence results for lots of known successful cases are still unclear;
- Gap from the stationary point to the global optimizer.

Thank you for your attention!

Email: liuxin@lsec.cc.ac.cn