Flexible ADMM for Block-Structured Convex and Nonconvex Optimization

Zhi-Quan (Tom) Luo

Joint work with Mingyi Hong, Tsung-Hui Chang, Xiangfeng Wang, Meisam Razaviyanyn, Shiqian Ma

University of Minnesota

September, 2014

Problem

- We consider the following block-structured problem

$$\begin{aligned} \text{minimize} \quad & f(x) := g(x_1, x_2, \cdots, x_K) + \sum_{k=1}^{K} h_k(x_k) \\ \text{subject to} \quad & Ex := E_1x_1 + E_2x_2 + \cdots + E_Kx_K = q \\ & x_k \in X_k, \quad k = 1, 2, ..., K, \end{aligned} \tag{1.1}$$

- $x := (x_1^T, ..., x_K^T)^T \in \Re^n$ is a partition of the optimization variable x, $X = \prod_{k=1}^{K} X_k$ is the feasible set for x
- $g(\cdot)$: smooth, possibly nonconvex; coupling all variables
- $h_k(\cdot)$: convex, possibly nonsmooth
- $E := (E_1, E_2, ..., E_K) \in \Re^{m \times n}$ is a partition of E

Applications

Lots of emerging applications

- **Compressive Sensing** Estimate a sparse vector x by solving the following ($K = 2$) [Candes 08]:

$$\begin{array}{ll} \text{minimize} & \|z\|^2 + \lambda\|x\|_1 \\ \text{subject to} & Ex + z = q, \end{array}$$

  where E is a (fat) observation matrix and $q \approx Ex$ is a noisy observation vector

- If we require $x \geq 0$ then we obtain a three block ($K = 3$) convex separable optimization problem

Applications (cont.)

- **Stable Robust PCA** Given a noise-corrupted observation matrix $M \in \Re^{m\times n}$, separate a low rank matrix L and a sparse matrix S [Zhou 10]

$$\begin{array}{ll} \text{minimize} & \|L\|_* + \rho\|S\|_1 + \lambda\|Z\|_F^2 \\ \text{subject to} & L + S + Z = M \end{array}$$

- $\|\cdot\|_*$: the matrix nuclear norm
- $\|\cdot\|_1$ and $\|\cdot\|_F$ denote the ℓ_1 and the Frobenius norm of a matrix
- Z denotes the noise matrix

Applications: The BP Problem

- Consider the basis pursuit (BP) problem [Chen et al 98]

$$\min_{x} \|x\|_1 \quad \text{s.t.} \quad Ex = q, \; x \in X.$$

- Partition x by $x = [x_1^T, \cdots, x_K^T]^T$ where $x_k \in \Re^{n_k}$
- Partition E accordingly
- The BP problem becomes a K block problem

$$\min_{x} \sum_{k=1}^{K} \|x_k\|_1 \quad \text{s.t.} \quad \sum_{k=1}^{K} E_k x_k = q, \; x_k \in X_k, \; \forall\, k.$$

Applications: Wireless Networking

- Consider a network with K secondary users (SUs), L primary users (PUs) and a secondary BS (SBS)
- s_k: user k's transmit power; r_k the channel between user k and the SBS; P_k SU k's total power budget
- $g_{k\ell}$: the channel between the kth SU to the ℓth PU

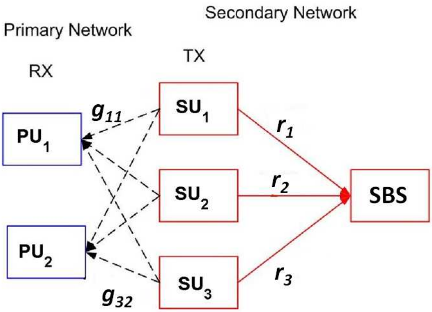

Figure: Illustration of the CR network.

Applications: Wireless Networking

- **Objective** maximize the SUs' throughput, subject to limited interference to PUs:

$$\begin{aligned} \max_{\{s_k\}} \quad & \log\left(1+\sum_{k=1}^{K}|r_k|^2 s_k\right) \\ \text{s.t.} \quad & 0 \le \; s_k \le P_k, \; \sum_{k=1}^{K}|g_{k\ell}|^2 s_k \le I_\ell, \; \forall\, \ell, \; k, \end{aligned}$$

- Again in the form of (1.1)
- Similar formulation for systems with multiple channels, multiple transmit/receive antennas

Application: DR in Smart Grid Systems

- Utility company bids the electricity from the power market
- Total cost

  Bidding cost in a wholesale day-ahead market
  Bidding cost in real-time market
- The demand response (DR) problem [Alizadeh et al 12]

  Utility have control over the power consumption of users' appliances (e.g., controlling the charging rate of electrical vehicles)

  Objective: minimize the total cost

Application: DR in Smart Grid Systems

- K customers, L periods
- $\{p_\ell\}_{\ell=1}^{L}$: the bids in a day-ahead market for a period L
- $\mathbf{x}_k \in \Re^{n_k}$: control variables for the appliances of customer k
- **Objective**: Minimize the bidding cost + power imbalance cost, by optimizing the bids and controlling the appliances [Chang et al 12]

$$\min_{\{\mathbf{x}_k\},\mathbf{p},\mathbf{z}} \quad C_p(\mathbf{z}) + C_s\Big(\mathbf{z} + \mathbf{p} - \sum_{k=1}^{K} \mathbf{\Psi}_k \mathbf{x}_k\Big) + C_d(\mathbf{p})$$

$$\text{s.t.} \quad \sum_{k=1}^{K} \mathbf{\Psi}_k \mathbf{x}_k - \mathbf{p} - \mathbf{z} \le 0, \ \mathbf{z} \ge 0, \ \mathbf{p} \ge 0, \ \mathbf{x}_k \in X_k, \ \forall\, k.$$

Challenges

- For huge scale (BIG data) applications, efficient algorithms needed
- Many existing first-order algorithms do not apply
  - The block coordinate descent algorithm (BCD) cannot deal with linear coupling constraints [Bertsekas 99]
  - The block successive upper-bound minimization (BSUM) method cannot apply either [Razaviyayn-Hong-Luo 13]
  - The alternating direction method of multipliers (ADMM) only works for convex problem with 2 blocks of variables and separable objective [Boyd *et al* 11][Chen *et al* 13]
- General purpose algorithms can be very slow

Agenda

- The ADMM for multi-block structured convex optimization
  - The main steps of the algorithm
  - Rate of convergence analysis
- The BSUM-M for multi-block structured convex optimization
  - The main steps of the algorithm
  - Convergence analysis
- The flexible ADMM for structured nonconvex optimization
  - The main steps of the algorithm
  - Convergence analysis
- Conclusions

Agenda

- The ADMM for multi-block structured convex optimization
  - The main steps of the algorithm
  - Rate of convergence analysis
- The BSUM-M for multi-block structured convex optimization
  - The main steps of the algorithm
  - Convergence analysis
- The flexible ADMM for structured nonconvex optimization
  - The main steps of the algorithm
  - Convergence analysis
- Conclusions

The ADMM Algorithm

- The augmented Lagrangian function for problem (1.1) is

$$L(x;y) = f(x) + \langle y, q - Ex \rangle + \frac{\rho}{2}\|q - Ex\|^2, \tag{1.2}$$

  where $\rho \geq 0$ is a constant
- The primal problem is given by

$$d(y) = \min_x \; f(x) + \langle y, q - Ex \rangle + \frac{\rho}{2}\|q - Ex\|^2 \tag{1.3}$$

- The dual problem is

$$d^* = \max_y d(y), \tag{1.4}$$

  d^* equals to the optimal solution of (1.1) under mild conditions

The ADMM Algorithm

Alternating Direction Method of Multipliers (ADMM)

At each iteration $r \geq 1$, first update the primal variable blocks in the Gauss-Seidel fashion and then update the dual multiplier:

$$\begin{cases} x_k^{r+1} = \arg\min_{x_k \in X_k} L(x_1^{r+1}, ..., x_{k-1}^{r+1}, x_k, x_{k+1}^r, ..., x_K^r; y^r), \ \forall\ k \\ y^{r+1} = y^r + \alpha(q - Ex^{r+1}) = y^r + \alpha\left(q - \sum_{k=1}^{K} E_k x_k^{r+1}\right), \end{cases}$$

where $\alpha > 0$ is the step size for the dual update.

- Inexact primal minimization $\Rightarrow q - Ex^{t+1}$ is no longer the dual gradient!
- Dual ascent property $d(y^{t+1}) \geq d(y^t)$ is lost
- Consider $\alpha = 0$, or $\alpha \approx 0$...

The ADMM Algorithm (cont.)

- The Alternating Direction Method of Multipliers (ADMM) optimizes the augmented Lagrangian function one block variable at each time [Boyd 11, Bertsekas 10]
- Recently found lots of applications in large-scale structured optimization; see [Boyd 11] for a survey
- Highly efficient, especially when the per-block subproblems are easy to solve (with closed-form solution)
- Used widely (*wildly?*), even to nonconvex problems, with no guarantee of convergence

Known Convergence Results and Challenges

- $K = 1$: reduces to the conventional dual ascent algorithm [Bertsekas 10]; The convergence and rate of convergence has been analyzed in [Luo 93, Tseng 87]
- $K = 2$: a special case of Douglas-Rachford splitting method, and its convergence is studied in [Douglas 56, Eckstein 89]
- $K = 2$: the rate of convergence has recently been studied in [Deng 12]; analysis based on strong convexity and a contraction argument; Iteration complexity has been studied in [He 12]

Main Challenges: How about $K \geq 3$?

- Oddly, when $K \geq 3$, there is little convergence analysis
- Recently [Chen *et al* 13] discovered a counter example showing three-block ADMM is not necessarily convergent
- When $f(\cdot)$ is strongly convex, and when α is small enough, the algorithm converges [Han-Yuan 13]
- Some relaxed condition has been given recently in [Lin-Ma-Zhang 14], but still need $K-1$ blocks to be strongly convex
- What about the case when $f_k(\cdot)$'s are convex but not strongly convex? nonsmooth?
- Besides convergence, can we characterize how fast the algorithm converges?

Agenda

- The ADMM for multi-block structured convex optimization
  - The main steps of the algorithm
  - Rate of convergence analysis
- The BSUM-M for multi-block structured convex optimization
  - The main steps of the algorithm
  - Convergence analysis
- The flexible ADMM for structured nonconvex optimization
  - The main steps of the algorithm
  - Convergence analysis
- Conclusions

Our Main Result [Hong-Luo 12]

Suppose some regularity conditions hold. If the stepsize α is sufficiently small, then

- the sequence of iterates $\{(x^r, y^r)\}$ generated by the ADMM algorithm (12) converges linearly to an optimal primal-dual solution for (1.1).
- the sequence of feasibility violation $\{\|Ex^r - q\|\}$ converges linearly.

- No strong convexity assumed
- Linear convergence here means certain measure of optimality gap shrinks by a constant factor after each ADMM iteration
- This result applies to any finite $K > 0$

Main Assumptions

The following are the main assumptions regarding f:

(a) The global minimum of (1.1) is attained and so is its dual optimal value

(b) The smooth part g further decomposable as

$$g(x_1, \cdots, x_k) = \sum_{k=1}^{K} g_k(A_k x_k)$$

where g_k is convex; A_k's are some given matrices (not necessarily full column rank)

(c) Each g_k is strictly convex and continuously differentiable with a uniform Lipschitz continuous gradient

$$\|A_k^T \nabla g_k(Ax_k) - A_k^T \nabla g_k(Ax_k')\| \leq L\|x_k - x_k'\|, \ \forall\ x_k, x_k' \in X_k$$

Main Assumptions (cont.)

(d) Each h_k satisfies either one of the following conditions

(1) The epigraph of $h_k(x_k)$ is a polyhedral set.

(2) $h_k(x_k) = \lambda_k \|x_k\|_1 + \sum_J w_J \|x_{k,J}\|_2$, where $x_k = (\cdots, x_{k,J}, \cdots)$ is a partition of x_k with J being the partition index.

(3) Each $h_k(x_k)$ is the sum of the functions described in the previous two items.

(e) Each submatrix E_k has full column rank.

(f) The feasible sets X_k's are compact polyhedral sets.

Preliminary: Measures of Optimality (cont.)

- Let $X(y^r)$ denote the set of optimal solutions for

$$d(y^r) = \min_x L(x; y^r),$$

and let

$$\bar{x}^r = \operatorname*{argmin}_{\bar{x} \in X(y^r)} \|\bar{x} - x^r\|.$$

- Let us define

$$\text{dist}\,(x^r, X(y^r)) = \min_{\bar{x} \in X(y^r)} \|\bar{x} - x^r\|,$$

and

$$\text{dist}\,(y^r, Y^*) = \min_{\bar{y} \in Y^*} \|\bar{y} - y^r\|.$$

The Key Idea

- Define the dual optimality gap as

$$\Delta_d^r = d^* - d(y^r) \geq 0.$$

- Define the primal optimality gap as

$$\Delta_p^r = L(x^{r+1}; y^r) - d(y^r) \geq 0.$$

- If $\Delta_d^r + \Delta_p^r = 0$, then an optimal solution is obtained
- **The Key Step**: Show that the combined dual and primal gaps $\Delta_d^r + \Delta_p^r$ decreases linearly in each iteration

Illustration of the Gaps (iteration r)

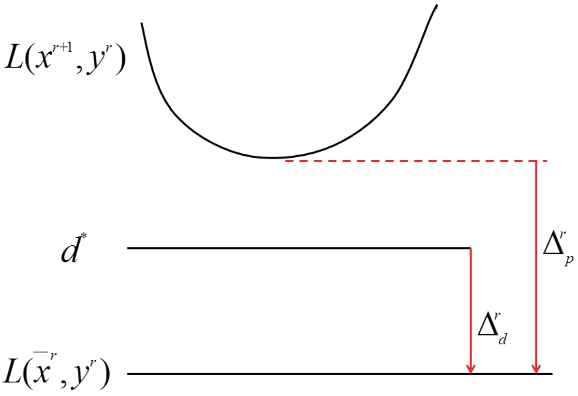

Figure: Illustration of the reduction of the combined gap.

Illustration of the Gaps (iteration $r + 1$)

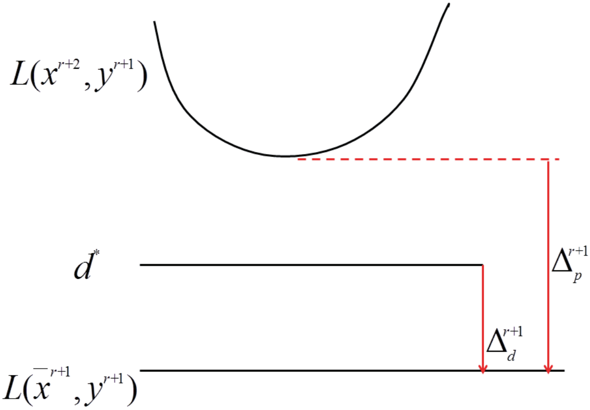

Figure: Illustration of the reduction of the combined gap.

Illustration of the Gaps (iteration $r+2$)

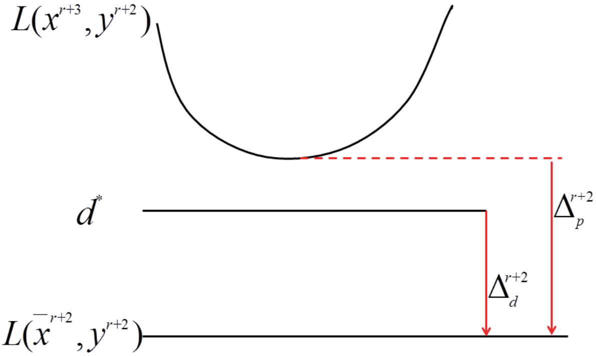

Figure: Illustration of the reduction of the combined gap.

Agenda

- The ADMM for multi-block structured convex optimization
  - The main steps of the algorithm
  - Rate of convergence analysis
- The BSUM-M for multi-block structured convex optimization
  - The main steps of the algorithm
  - Convergence analysis
- The flexible ADMM for structured nonconvex optimization
  - The main steps of the algorithm
  - Convergence analysis
- Conclusions

The BSUM-M Algorithm: Motivation and Main Ideas

- **Questions**
  - Can we do inexact primal update (i.e., proximal update)?
  - How to choose the dual stepsize α?
  - Can we consider more flexible block selection rules?
- To address these questions, we introduce the Block Successive Upperbound Minimization method of Multipliers (BSUM-M)
- **Main idea: Primal update**

  Pick the primal variables either sequentially or randomly
  Optimize some approximate version of $L(x, y)$
- **Main idea: Dual update**

  Inexact dual ascent + proper step size control

The BSUM-M Algorithm: Details

- At iteration $r+1$, a block variable x_k is updated by solving

$$\min_{x_k \in X_k} \quad u_k\left(x_k; x_1^{r+1}, \cdots, x_{k-1}^{r+1}, x_k^r, \cdots, x_K^r\right) + \langle y^{r+1}, q - E_k x_k\rangle + h_k(x_k)$$

- $u_k(\cdot\,;\, x_1^{r+1}, \cdots, x_{k-1}^{r+1}, x_k^r, \cdots, x_K^r)$: is an *upper-bound* of

$$g(x) + \frac{\rho}{2}\|q - Ex\|^2$$

  at the current iterate $(x_1^{r+1}, \cdots, x_{k-1}^{r+1}, x_k^r, \cdots, x_K^r)$
- Proximal gradient step, proximal point step are special cases

The BSUM-M Algorithm: G-S Update Rule

The BSUM-M Algorithm

At each iteration $r \geq 1$:

$$\begin{cases} y^{r+1} = y^r + \alpha^r(q - Ex^r) = y^r + \alpha^r\left(q - \sum_{k=1}^{K} E_k x_k^r\right), \\ x_k^{r+1} = \arg\min_{x_k \in X_k} u_k(x_k; w_k^{r+1}) - \langle y^{r+1}, E_k x_k \rangle + h_k(x_k), \ \forall\ k \end{cases}$$

where $\alpha^r > 0$ is the dual stepsize.

- To simplify notations, we have defined

$$w_k^{r+1} := (x_1^{r+1}, \cdots, x_{k-1}^{r+1}, x_k^r, x_{k+1}^r, \cdots, x_K^r),$$

The BSUM-M Algorithm: Randomized Update Rule

- Select a vector $\{p_k > 0\}_{k=0}^K$ such that $\sum_{k=0}^K p_k = 1$
- Each iteration "t" only updates a single randomly selected primal or dual variable

The Randomized BSUM-M Algorithm

At iteration $t \geq 1$, pick $k \in \{0, \cdots, K\}$ with probability p_k and

If $k = 0$

$$y^{t+1} = y^t + \alpha^t(q - Ex^t),$$

$$x_k^{t+1} = x_k^t, \; k = 1, \cdots, K.$$

Else If $k \in \{1, \cdots, K\}$

$$x_k^{t+1} = \operatorname{argmin}_{x_k \in X_k} u_k(x_k; x^t) - \langle y^r, E_k x_k \rangle + h_k(x_k),$$

$$x_j^{t+1} = x_j^t, \; \forall\, j \neq k, \quad y^{t+1} = y^t.$$

End

Key Features

- Primal update similar to (randomized) BCD [Nestrov 12] [Richtárik- Takáč12] [Saha-Tewari 13]; but can deal with linear coupling constraint
- Primal-dual update similar to ADMM; but can deal with multiple coupled blocks
- Using approximate upper bound function – closed-form subproblem
- Flexibility in update schedule – deterministic+randomized
- **Key Questions**

  How to select the approximate upper bound function
  How to select the primal/dual stepsize (ρ, α)
  Guaranteed convergence?

Convergence Analysis: Assumptions

- **Assumption A** (on the problem)

(a) Problem (1.1) is convex and feasible

(b) $g(x) = \ell(Ax) + \langle x, b\rangle$; $\ell(\cdot)$ smooth strictly convex, A not necessarily full column rank

(c) Nonsmooth function h_k:

$$h_k(x_k) = \lambda_k \|x_k\|_1 + \sum_J w_J \|x_{k,J}\|_2,$$

where $x_k = (\cdots, x_{k,J}, \cdots)$ is a partition of x_k; $\lambda_k \geq 0$ and $w_J \geq 0$ are some constants.

(d) The feasible sets $\{X_k\}$ are compact polyhedral sets, and are given by $X_k := \{x_k \mid C_k x_k \leq c_k\}$.

Convergence Analysis: Assumptions

- **Assumption B** (on u_k)

(a) $u_k(v_k;x) \geq g(v_k, x_{-k}) + \frac{\rho}{2}\|E_k v_k - q + E_{-k}x_{-k}\|^2, \quad \forall\, v_k \in X_k,\ \forall\, x, k$ (upper-bound)

(b) $u_k(x_k;x) = g(x) + \frac{\rho}{2}\|Ex - q\|^2,\ \forall\, x, k$ (locally tight)

(c) $\nabla u_k(x_k;x) = \nabla_k\left(g(x) + \frac{\rho}{2}\|Ex - q\|^2\right),\ \forall\, x, k$

(d) For any given x, $u_k(v_k;x)$ is strongly convex in v_k

(e) For given x, $u_k(v_k;x)$ has Lipchitz continuous gradient

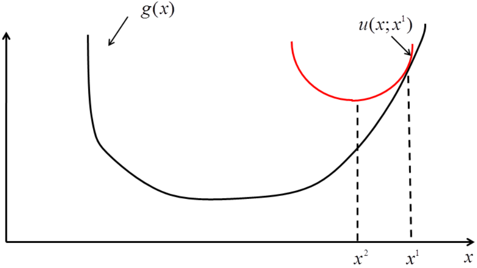

Figure: Illustration of the upper-bound.

The Convergence Result [Hong *et al* 13]

Suppose Assumptions A-B hold, and the dual stepsize α^r satisfies

$$\sum_{r=1}^{\infty} \alpha^r = \infty, \quad \lim_{r\to\infty} \alpha^r = 0.$$

Then we have the following:

- For the BSUM-M, we have $\lim_{r\to\infty} \|Ex^r - q\| = 0$, and every limit point of $\{x^r, y^r\}$ is a primal and dual optimal solution.
- For the RBSUM-M, we have $\lim_{t\to\infty} \|Ex^t - q\| = 0$ w.p.1. Further, every limit point of $\{x^t, y^t\}$ is a primal and dual optimal solution w.p.1.

Numerical Result: Counterexample for multi-block ADMM

- Recently [Chen-He-Ye-Yuan 13] shows (through an example) that applying ADMM to multi-block problem can diverge
- We show applying (R)BSUM-M to the same problem converges
- **Main message**: Dual stepsize control is crucial
- Consider the following linear systems of equations (unique solution $x_1 = x_2 = x_3 = 0$)

$$E_1 x_1 + E_2 x_2 + E_3 x_3 = 0,$$

$$\text{with} \quad [E_1 \ E_2 \ E_3] = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{bmatrix}.$$

Counterexample for multi-block ADMM (cont.)

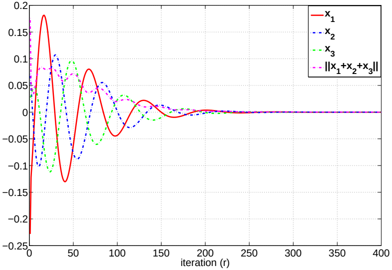

Figure: Iterates generated by the BSUM-M. Each curve is averaged over 1000 runs (with random starting points).

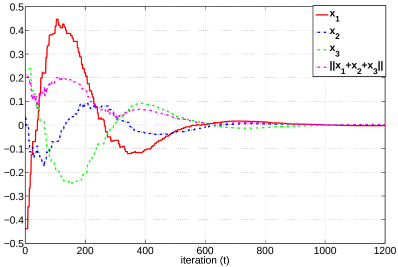

Figure: Iterates generated by the RBSUM-M algorithm. Each curve is averaged over 1000 runs (with random starting points)

Agenda

- The ADMM for multi-block structured convex optimization
  - The main steps of the algorithm
  - Rate of convergence analysis
- The BSUM-M for multi-block structured convex optimization
  - The main steps of the algorithm
  - Convergence analysis
- The flexible ADMM for structured nonconvex optimization
  - The main steps of the algorithm
  - Convergence analysis
- Conclusions

ADMM for nonconvex problem?

- ADMM is known to work for separable convex problems
- But ADMM is also known to work well for nonconvex problems, at least empirically
  - Nonnegative matrix factorization [Zhang 10] [Sun-Fevotte 14]
  - Phase retrieval [Wen *et al* 12]
  - Distributed matrix factorization [Ling-Xu-Yin-Wen 12]
  - Polynomial optimization [Jiang-Ma-Zhang 13]
  - Asset allocation [Wen *et al* 13]
  - Zero variance discriminant analysis [Ames-Hong 14]
  - ...
- Although ADMM works very well empirically, theoretically little is known
- To show convergence, most of the analysis assumes favorable properties on the iterates generated by the algorithm...

Convergence analysis of ADMM for nonconvex problems

- It is indeed possible to show ADMM globally converges for nonconvex problems [Hong-Luo 14]
  - For a family of nonconvex consensus problems
  - For a family of nonconvex, multi-block sharing problems
- Key ingredients:
  - Consider the vanilla ADMM
  - Keep primal and dual stepsize identical ($\alpha = \rho$)
  - ρ large enough to make each subproblem strongly convex
  - Use the augmented Lagrangian as the potential function
- Our analysis can extend to flexible block selection rules
  - Gauss-Seidel block selection rule
  - Randomized block selection rule
  - Essentially Cyclic block selection rule

The Consensus Problem

- Consider the following nonconvex problem

$$\begin{aligned} \min \quad & f(x) := \sum_{k=1}^{K} g_k(x) + h(x) \\ \text{s.t.} \quad & x \in X \end{aligned} \tag{3.5}$$

- g_k: smooth, possibly nonconvex functions
- h: is a convex nonsmooth regularization term
- This is the global consensus problem discussed heavily in [Section 7, Boyd *et al* 11], but there only convex cases are considered

The Consensus Problem (cont.)

- In some applications, each g_k handled by a single agent
- This motivates the following consensus formulation

$$\begin{aligned} \min \quad & \sum_{k=1}^{K} g_k(x_k) + h(x) \\ \text{s.t.} \quad & x_k = x, \ \forall\ k = 1, \cdots, K, \quad x \in X. \end{aligned} \tag{3.6}$$

- The augmented Lagrangian is given by

$$\begin{aligned} L(\{x_k\}, x; y) = & \sum_{k=1}^{K} g_k(x_k) + h(x) + \sum_{k=1}^{K} \langle y_k, x_k - x \rangle \\ & + \sum_{k=1}^{K} \frac{\rho_k}{2} \|x_k - x\|^2. \end{aligned}$$

The ADMM for the Consensus Problem

Algorithm 1. ADMM for the Consensus Problem

At each iteration $t+1$, compute:

$$x^{t+1} = \operatorname*{argmin}_{x \in X} L(\{x_k^t\}, x; y^t). \tag{3.7}$$

Each node k computes x_k by solving:

$$x_k^{t+1} = \arg\min_{x_k} g_k(x_k) + \langle y_k^t, x_k - x^{t+1} \rangle + \frac{\rho_k}{2} \|x_k - x^{t+1}\|^2. \tag{3.8}$$

Update the dual variable:

$$y_k^{t+1} = y_k^t + \rho_k \left(x_k^{t+1} - x^{t+1} \right). \tag{3.9}$$

Main Assumptions

Assumption C

C1. Each ∇g_k is Lipschitz Continuous with constant L_k; h is convex (possible nonsmooth)

C2. For all k, the stepsize ρ_k is chosen large enough such that:
- For all k, the x_k subproblem is strongly convex with modulus $\gamma_k(\rho_k)$;
- For all k, $\rho_k > \max\{\frac{2L_k^2}{\gamma_k(\rho_k)}, L_k\}$.

C3. $f(x)$ is lower bounded for all $x \in X$.

Convergence Analysis [Hong-Luo 14]

Suppose Assumption C is satisfied. Then

$$\lim_{t\to\infty} \|x_k^{t+1} - x^{t+1}\| = 0.$$

Further, we have the following

- Any limit point of the sequence generated by the ADMM is a stationary solution of problem (3.6).
- If X is a compact set, then the sequence converges to the set of stationary solutions of problem (3.6).

- Primal feasibility always satisfied in the limit
- No assumptions made on the iterates

The Sharing Problem

- Consider the following problem

$$\begin{aligned} \min \quad & f(x_1, \cdots, x_K) := \sum_{k=1}^{K} g_k(x_k) + \ell\left(\sum_{k=1}^{K} A_k x_k\right) \\ \text{s.t.} \quad & x_k \in X_k,\ k = 1, \cdots, K. \end{aligned} \tag{3.10}$$

- ℓ: smooth nonconvex
- g_k: either smooth nonconvex or convex (possibly nonsmooth)
- Similar to the well-known sharing problem discussed in [Section 7.3, Boyd *et al* 11], but allows nonconvex objective

Reformulation

- This problem can be equivalently formulated into

$$\begin{aligned} \min \quad & \sum_{k=1}^{K} g_k(x_k) + \ell(x) \\ \text{s.t.} \quad & \sum_{k=1}^{K} A_k x_k = x, \quad x_k \in X_k,\ k = 1, \cdots, K. \end{aligned} \tag{3.11}$$

- A K-block, nonconvex reformulation
- Even if g_k's and ℓ are convex, not clear whether ADMM converges

Main Assumptions

Assumption D

D1. $\nabla \ell(x)$ is Lipcshitz continuous with constant L; Each A_k full column rank, with $\rho_{\min}(A_k^T A_k) > 0$.

D2. The stepsize ρ is chosen large enough such that:

(1) each x_k and x subproblem is strongly convex, with modulus $\{\gamma_k(\rho)\}_{k=1}^K$ and $\gamma(\rho)$, respectively.

(2) $\rho > \max\left\{\frac{2L^2}{\gamma(\rho)}, L\right\}$.

D3. $f(x_1, \cdots, x_K)$ is lower bounded for all $x_k \in X_k$ and all k.

D4. g_k is either nonconvex Lipcshitz continuous with constant L_k, or convex (possibly nonsmooth).

Convergence Analysis [Hong-Luo 14]

Suppose Assumption D is satisfied. Then

$$\lim_{t\to\infty} \|x_k^{t+1} - x^{t+1}\| = 0.$$

Further, we have the following

- Every limit point generated by ADMM is a stationary solution of problem (3.11).
- If X_k is a compact set for all k, then ADMM converges to the set of stationary solutions of problem (3.11).

- Primal feasibility always satisfied in the limit
- No assumptions made on the iterates

Remarks

- For the sharing problem, if all objectives are convex, our result shows that multi-block ADMM converges with $\rho \geq \sqrt{2}L$
- Similar analysis applies for the 2-block reformulation of the sharing problem
- Analysis can be extended to include proximal block updates
- Analysis can be generalized to flexible block update rules – all x_k's do not need to update at the same time

Conclusions and Future Works

- We have shown the convergence and the rate of convergence for multiblock ADMM without strong convexity
- The key is to use the combined primal-dual gap as the potential function
- We introduce a new algorithm called BSUM-M that can solve multi-block linearly constrained convex problems
- The key is to use a diminishing dual stepsize
- We show that ADMM converges for two families of nonconvex, possibly multiple problems
- The key is to use the Augemented Lagrangian as the potential function

Conclusions and Future Works (cont.)

- Iteration complexity analysis for multi-block and/or nonconvex ADMM?
- Can we generalize the analysis for nonconvex ADMM to a wider range of problems?
- Nonlinearly constrained problems?

Thank You!

Reference

1 [Ames-Hong 14] Ames, B. and Hong, M. "Alternating directions method of multipliers for l1- penalized zero variance discriminant analysis and principal component analysis," Preprint

2 [Bertsekas 99] Bertsekas, D.P.: Nonlinear Programming. Athena Scientific.

3 [Boyd et al 11] Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J.: Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Foundations and Trends in Machine Learning.

4 [Candes 09] Candes, E and Plan , Y.: Ann. Statist.

5 [Chen-He-Ye-Yuan 13] C. Chen, B. He, X. Yuan, and Y. Ye, "The direct extension of admm for multi-block convex minimization problems is not necessarily convergent," 2013.

6 [Douglas 56] Douglas, J. and Rachford, H.H.: On the numerical solution of the heat conduction problem in 2 and 3 space variables. Trans. of the American Math. Soc.

Reference

7 [Deng 12] Deng W. and Yin. W.: On the global and linear convergence of the generalized alternating direction method of multipliers. Rice CAAM tech report

8 [Eckstein 89] Eckstein, J.: Splitting methods for monotone operators with applications to parallel optimization. Ph.D Thesis, Operations Research Center, MIT.

9 [Nestrov 12] Y. Nesterov, "Efficiency of coordiate descent methods on huge-scale optimization problems," SIAM Journal on Optimization, vol. 22, no. 2, 2012.

10 [Han-Yuan 12] Han D. and Yuan X.: A Note on the Alternating Direction Method of Multipliers, J Optim Theory Appl

11 [He-Yuan 12] He, B. S. and Yuan, X. M.: On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numer. Anal.

12 [Hong-Luo 12] Hong, M. and Z.-Q., Luo: On the linear convergence of ADMM Algorithm. Manuscript.

13 [Hong-Luo 14] Hong, M. and Z.-Q., Luo: Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems. Manuscript.

Reference

14 [Hong et al 13] Hong, M. et al: A Block Successive Upper Bound Minimization Method of Multipliers for Linearly Constrained Convex Optimization. Manuscript.

15 [Jiang-Ma-Zhang 13] Jiang, B. and Ma, S. and Zhang, S. "Alternating direction method of multipliers for real and complex polynomial optimization models," manuscript

16 [Lin-Ma-Zhang 14] Lin, T. and Ma, S. and Zhang, S. "On the Convergence Rate of Multi-Block ADMM," manuscript, 2014

17 [Ling et al 21] Ling, Q. et al, "Decentralized low-rank matrix completion," ICASSP, 2012

18 [Luo 93] Luo, Z.-Q. and Tseng, P.: On the convergence rate of dual ascent methods for strictly convex minimization. Math. of Oper. Res.

19 [Razaviyayn-Hong-Luo 13] Razaviyayn, M., and Hong, M. and Luo, Z.-Q.: A unified convergence analysis of block successive minimization methods for nonsmooth optimization. SIAM J. Opt. 2013

20 [Richtárik- Takáč12] P. Richtarik and M. Takac, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function," Mathematical Programming, 2012.

Reference

21 [Saha-Tewari 13] A. Saha and A. Tewari, "On the nonaymptotic convergence of cyclic coordinate descent method," SIAM Journal on Optimization, vol. 23, no. 1, 2013.

22 [Tseng 87] Tseng, P., and Bertsekas D. P.: Relaxation methods for problems with strictly convex separable costs and linear constraints. Math. Prog.

23 [Wang 13] Wang X. Hong M. Ma S. and Z.-Q. Luo: Solving Multiple-Block Separable Convex Minimization Problems Using Two-Block Alternating Direction Method of Multipliers. Manuscript

24 [Wen et al 12] Wen, Z. et al, "Alternating direction methods for classical and ptychographic phase retrieval," Inverse Problems, 2012.

25 [Yang 11] Yang J. and Zhang Y. Alternating direction algorithms for l1-problems in compressive sensing. SIAM J. on Scientific Comp.

26 [Zhou 10] Zhou, Z., Li, X., Wright, J., Candes, E.J., and Ma, Y.: Stable principal component pursuit. Proceedings of IEEE ISIT