Algorithms for Non-Convex Decentralized Optimization (b)

Mingyi Hong

University Of Minnesota

Outline

- ADMM algorithm for star network
- Primal-Dual algorithm for general connected networks
- Discussions / Recent advances

Unconstrained multiagent-optimization problem

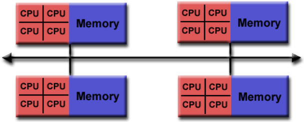
$$\begin{aligned} \text{minimize}_x \quad & f(x) := \sum_{i=1}^{m} g_i(x) \\ \text{subject to} \quad & x \in X \subseteq \mathbb{R}^n \end{aligned} \tag{1.1}$$

- In this section, we consider problems where each $g_i : \mathbb{R}^n \to \mathbb{R}$ is a non-convex function, which is known only to agent i.

The Star Network

A special case: The star network

- Consider a special case where there is a master node and a number of slave nodes
- The setting, although very simple, is already very popular in practice



The Global Consensus Problem

- Consider a nonconvex global consensus problem

$$\min \quad f(x) := \sum_{i=1}^{m} g_i(x) + h(x) \quad \text{subject to} \quad x \in X \qquad (2.1)$$

- $g_k(x)$: smooth, possibly nonconvex function

- $h(x)$: convex nonsmooth regularization term, or just $\equiv 0$

The Global Consensus Problem

- Each g_i needs to be handled by a single agent

- So we can reformulate the problem as below

$$\begin{aligned} \min \sum_{i=1}^{m} g_i(x_i) + h(x) \\ \text{subject to} \quad x_i = x, \; \forall \; i = 1, \cdots, m, \quad x \in X. \end{aligned} \tag{2.2}$$

- Here $h(\cdot)$ is the (local) objective function on the central controller / master node

- If you view the mater node as an agent, then $h(\cdot)$ can also be viewed as a local objective function

Application: Distributed Sparse-PCA

- The sparse PCA problem can be formulated as

$$\min_x x^T Bx + \gamma\|x\|_0,\ \|x\|_2^2 \le 1 \tag{2.3}$$

  where $-B \succ 0$

- Consider its relaxation

$$\min_x x^T Bx + \gamma\|x\|_1,\ \|x\|_2^2 \le 1 \tag{2.4}$$

- Has wide applications, for example large-scale text data analysis

Application: Distributed Sparse-PCA (cont.)

- In large-scale text analysis, let $C \in \mathbb{R}^{R \times N}$ denote the summary of the text corpora such that
  - A total of R documents (one row one document); N words
  - $C[r, n] = 1$ means word n appears in document r

Application: Distributed Sparse-PCA (cont.)

- In large-scale text analysis, let $C \in \mathbb{R}^{R\times N}$ denote the summary of the text corpora such that
  - A total of R documents (one row one document); N words
  - $C[r,n] = 1$ means word n appears in document r
- Standard sparse PCA analysis

$$\min_x -x^T C^T C x + \gamma \|x\|_1, \text{ s.t. } \|x\|_2^2 \le 1 \tag{2.5}$$

Application: Distributed Sparse-PCA (cont.)

- In large-scale text analysis, let $C \in \mathbb{R}^{R \times N}$ denote the summary of the text corpora such that
  - A total of R documents (one row one document); N words
  - $C[r, n] = 1$ means word n appears in document r
- Standard sparse PCA analysis

$$\min_x -x^T C^T C x + \gamma \|x\|_1, \text{ s.t. } \|x\|_2^2 \leq 1 \tag{2.5}$$

- Suppose data is divided by $C = [C_1^T, \cdots, C_m^T]^T$, where each $C_k \in \mathbb{R}^{r_i \times N}$ consists of a subset of documents

Application: Distributed Sparse-PCA (cont.)

- In large-scale text analysis, let $C \in \mathbb{R}^{R\times N}$ denote the summary of the text corpora such that
  - A total of R documents (one row one document); N words
  - $C[r,n]=1$ means word n appears in document r
- Standard sparse PCA analysis

$$\min_x -x^T C^T C x + \gamma\|x\|_1, \text{ s.t. } \|x\|_2^2 \le 1 \tag{2.5}$$

- Suppose data is divided by $C = [C_1^T, \cdots, C_m^T]^T$, where each $C_k \in \mathbb{R}^{r_i \times N}$ consists of a subset of documents
- C_k's stored on K different agents; no data exchange allowed

$$\min_x -\sum_{i=1}^m x^T C_i^T C_i x + \gamma\|x\|_1, \ \|x\|_2^2 \le 1 \tag{2.6}$$

Application: Distributed Sparse-PCA (cont.)

- In large-scale text analysis, let $C \in \mathbb{R}^{R \times N}$ denote the summary of the text corpora such that
  - A total of R documents (one row one document); N words
  - $C[r, n] = 1$ means word n appears in document r
- Standard sparse PCA analysis

$$\min_x -x^T C^T C x + \gamma \|x\|_1, \text{ s.t. } \|x\|_2^2 \leq 1 \tag{2.5}$$

- Suppose data is divided by $C = [C_1^T, \cdots, C_m^T]^T$, where each $C_k \in \mathbb{R}^{r_i \times N}$ consists of a subset of documents
- C_k's stored on K different agents; no data exchange allowed

$$\min_x -\sum_{i=1}^{m} x^T C_i^T C_i x + \gamma \|x\|_1, \ \|x\|_2^2 \leq 1 \tag{2.6}$$

- A nonconvex global consensus problem

The Algorithm

- The augmented Lagrangian function is given by

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

Algorithm 1. The Classical ADMM for the Consensus Problem (2.2)

At each iteration $t+1$, compute:

$$x^{t+1} = \underset{x \in X}{\operatorname{argmin}}\, L(\{x_i^t\}, x; y^t) = \mathsf{prox}_{\iota(X)+h} \left[\frac{\sum_{i=1}^m \rho_i x_i^t + \sum_{i=1}^m y_i^t}{\sum_{i=1}^m \rho_i} \right].$$

Each node i computes x_i by solving:

$$x_i^{t+1} = \arg\min_{x_i} g_k(x_i) + \langle y_i^t, x_i - x^{t+1} \rangle + \frac{\rho_i}{2} \|x_i - x^{t+1}\|^2.$$

Each node i updates the dual variable:

$$y_i^{t+1} = y_i^t + \rho_i \left(x_i^{t+1} - x^{t+1} \right).$$

Convergence?

- Does the ADMM algorithm converge?
- What do we know about convergence for ADMM? It typically works for convex problems (at least pre-2014)
- Why?

A Toy Example

- First consider the following toy nonconvex example

$$\min \quad \frac{1}{2}x^T Ax + bx, \quad \text{subject to} \quad x \in [1, 2]$$

where A is a symmetric matrix, and $x \in \mathbb{R}^n$

A Toy Example

- First consider the following toy nonconvex example

$$\min \quad \frac{1}{2}x^T Ax + bx, \quad \text{subject to} \quad x \in [1,2]$$

where A is a symmetric matrix, and $x \in \mathbb{R}^n$

- Consider the following reformulation

$$\min \quad \frac{1}{2}x^T Ax + bz, \quad \text{subject to} \quad z \in [1,2], \; z = x$$

where A is a symmetric matrix not necessarily PSD

A Toy Example

- First consider the following toy nonconvex example

$$\min \quad \frac{1}{2}x^T Ax + bx, \quad \text{subject to} \quad x \in [1, 2]$$

where A is a symmetric matrix, and $x \in \mathbb{R}^n$

- Consider the following reformulation

$$\min \quad \frac{1}{2}x^T Ax + bz, \quad \text{subject to} \quad z \in [1, 2],\ z = x$$

where A is a symmetric matrix not necessarily PSD

- Consider the ADMM iteration with the update sequence $x \to z \to y$

A Toy Example (cont.)

- Convergence?
- Randomly generate the data matrices A and b
- Plot the following
  1. Primal feasibility gap $\|z - x\|$
  2. The optimality measure $\|x - \text{proj}\,[x - (Ax + b)]\,\|$
  3. The x-feasibility gap $\|x - \text{proj}(x)\|$
- The algorithm converges to a KKT point iff all three quantities go to zero

A Toy Example (cont.)

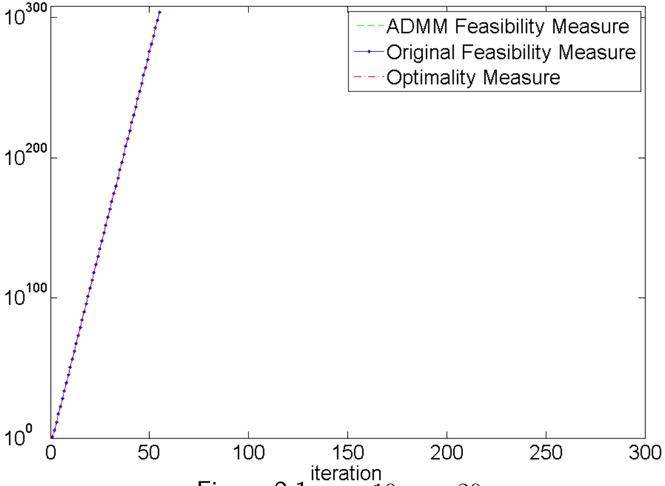


Figure 2.1: $n = 10$, $\rho = 20$

A Toy Example (cont.)

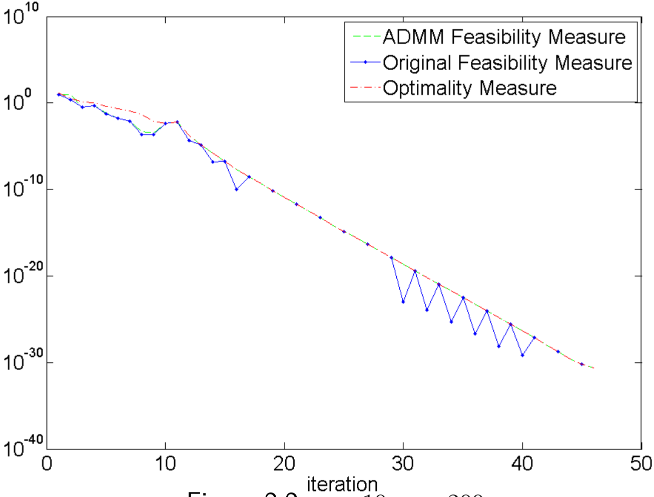


Figure 2.2: $n = 10$, $\rho = 200$

A Toy Example (cont.)

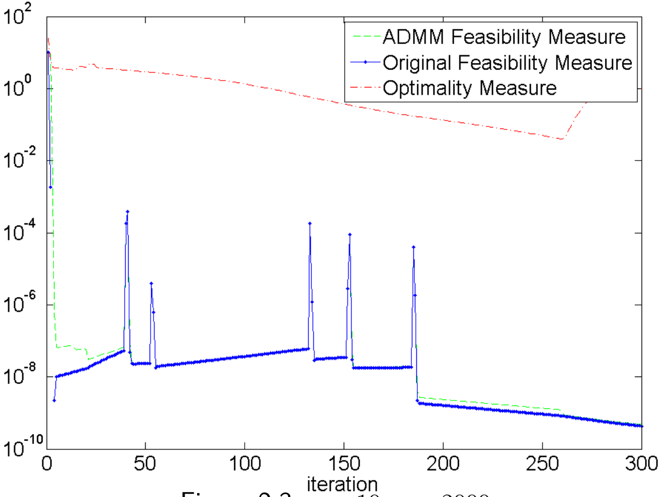


Figure 2.3: $n = 10$, $\rho = 2000$

A Toy Example (cont.)

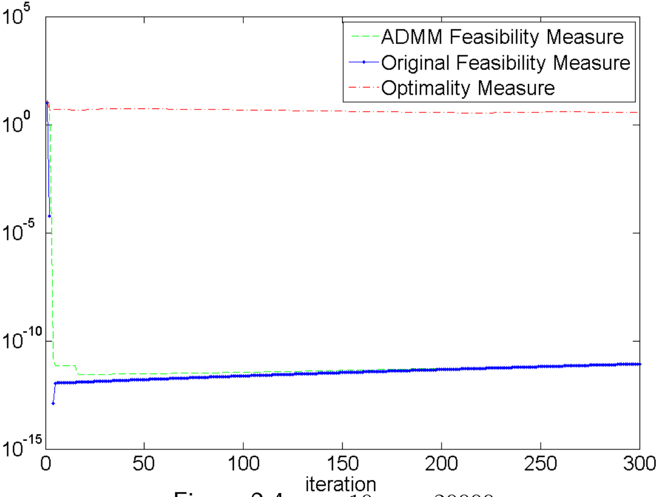


Figure 2.4: $n = 10$, $\rho = 20000$

A Toy Example (cont.)

- The convergence is ρ-dependent
- When ρ is small, the algorithm fails to converge
- When ρ is large, maybe convergent
- Different from the convex cases (where ρ does not affect convergence)

Assumptions

Assumption A.

A1. g_i's Lipschitz continuous:

$$\|\nabla_i g_i(x_i) - \nabla_i g_i(z_i)\| \leq L_i \|x_i - z_i\|, \ \forall\ x_i, z_i,\ i = 1, \cdots, m.$$

Moreover, h is convex (possible nonsmooth); X is a closed convex set.

A2. ρ_i is large enough such that:

1. For all i, the x_i subproblem is strongly convex with modulus $\gamma_i(\rho_i)$;
2. For all i, $\rho_i \gamma_i(\rho_i) > 2L_i^2$ and $\rho_i \geq L_i$.

A3. $f(x)$ is bounded from below over X.

Comments:

- As ρ_i inceases, x_i subproblem eventually becomes strongly convex
- By construction, the x subproblem is also strongly convex
- No assumption on the iterates generated by the algorithm

Proof Steps

- Use $L(x, \{x_i\}; y)$ as the potential function
- Use a three-step approach
- **Step 1**: Sufficient Descent

$$L(x^{t+1}, \{x_i^{t+1}\}; y^{t+1}) - L(x^t, \{x_i^t\}; y^t)$$
$$\leq -\sigma_0 \|x^{t+1} - x^t\|^2 - \sum_{i=1}^{m} \sigma_k \|x_i^{t+1} - x_i^t\|^2$$

- **Step 2**: Show that $L(x^{t+1}, \{x_i^{t+1}\}; y^{t+1})$ is lower bounded
- **Step 3**: Show that $\{x^{t+1}, \{x_i^{t+1}\}, y^{t+1}\}$ converges to a stationary solution of the global consensus problem
- The detailed proof follows from [H.-Luo-Razaviyayn 16][1]

[1] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," SIAM Journal On Optimization, 2016

S1: Sufficient Descent

- To show sufficient descent, we need the following lemma
- **Lemma**: The following is true

$$L_k^2\|x_i^{t+1} - x_i^t\|^2 \geq \|y_k^{t+1} - y_i^t\|^2, \; \forall\; i = 1, \cdots, m, \tag{2.7}$$

- From the x_i update step, we have the following optimality condition

$$\nabla g_i(x_i^{t+1}) + y_i^t + \rho_i(x_i^{t+1} - x^{t+1}) = 0, \; \forall\; i. \tag{2.8}$$

which implies

$$\nabla g_i(x_i^{t+1}) = -y_i^{t+1}, \; \forall\; i. \tag{2.9}$$

- Using the Lipschitz continuity assumption on ∇g_i we are done

S1: Sufficient Descent (cont.)

- Using the previous lemma, we can show the following

$$
\begin{aligned}
& L(\{x_i^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_i^t\}, x^t; y^t) \\
& \leq \sum_{i=1}^{m} \left(\frac{L_i^2}{\rho_i} - \frac{\gamma_i(\rho_i)}{2} \right) \|x_i^{t+1} - x_i^t\|^2 - \frac{\gamma}{2} \|x^{t+1} - x^t\|^2. \quad \text{(2.10)}
\end{aligned}
$$

  where $\gamma = \sum_{i=1}^{m} \rho_i$

- First split the successive difference of the augmented Lagrangian by

$$
\begin{aligned}
& L(\{x_i^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_i^t\}, x^t; y^t) \\
& = \left(L(\{x_i^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_i^{t+1}\}, x^{t+1}; y^t) \right) \\
& \quad + \left(L(\{x_i^{t+1}\}, x^{t+1}; y^t) - L(\{x_i^t\}, x^t; y^t) \right) \quad \text{(2.11)}
\end{aligned}
$$

S1: Sufficient Descent (cont.)

- The red term can be bounded by

$$
\begin{aligned}
& L(\{x_i^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_i^{t+1}\}, x^{t+1}; y^t) \\
& = \sum_{i=1}^{m} \langle y_i^{t+1} - y_i^t, x_i^{t+1} - x^{t+1} \rangle = \sum_i \rho_i \|x_i^{t+1} - x^{t+1}\|^2 \\
& = \sum_i \frac{1}{\rho_i} \|y_i^{t+1} - y_i^t\|^2 \le \sum_i \frac{L_i^2}{\rho_i} \|x_i^{t+1} - x_i^t\|^2 \qquad (2.12)
\end{aligned}
$$

- The blue term can be bounded by (using per-block strong convexity)

$$
\begin{aligned}
& L(\{x_i^{t+1}\}, x^{t+1}; y^t) - L(\{x_i^t\}, x^t; y^t) \\
& = L(\{x_i^{t+1}\}, x^{t+1}; y^t) - L(\{x^t{}_i\}, x^{t+1}; y^t) \\
& \quad + L(\{x^t{}_i\}, x^{t+1}; y^t) - L(\{x_k^t\}, x^t; y^t) \\
& \le -\sum_i \frac{\gamma_i(\rho_i)}{2} \|x_i^{t+1} - x_i^t\|^2 - \frac{\gamma}{2} \|x^{t+1} - x^t\|^2, \qquad (2.13)
\end{aligned}
$$

S2: Lower Bounds (skip)

- We then show that for some constant L^*

$$\lim_{t\to\infty} L(\{\{x_i^t\}, x^t, y^t\} = L^* > -\infty \tag{2.14}$$

- We have the following inequalities

$$
\begin{aligned}
& L(\{x_i^{t+1}\}, x^{t+1}; y^{t+1}) \\
& = h(x^{t+1}) + \sum_{i=1}^{m} g_i(x_i^{t+1}) + \langle y_i^{t+1}, x_i^{t+1} - x^{t+1}\rangle + \frac{\rho_i}{2}\|x_i^{t+1} - x^{t+1}\|^2 \\
& = h(x^{t+1}) + \sum_{i=1}^{m} g_i(x_i^{t+1}) + \langle \nabla g_i(x_i^{t+1}), x^{t+1} - x_i^{t+1}\rangle + \frac{\rho_k}{2}\|x_i^{t+1} - x^{t+1}\|^2 \\
& \overset{\text{(a)}}{\geq} h(x^{t+1}) + \sum_{i=1}^{m} g_i(x^{t+1}) = f(x^{t+1})
\end{aligned} \tag{2.15}
$$

where (a) comes from the Lipschitz continuity assumption, and the fact that $\rho_i \geq L_i$ for all $i = 1, \cdots, m$

- By assumption A3, $f(x)$ is lower bounded over $x \in X$

S3: Convergence

- **Theorem** Assume that Assumption A is satisfied. Then we have the following

1. We have $\lim_{t\to\infty} \|x_i^{t+1} - x^{t+1}\| = 0,\ k = 1, \cdots, K$
2. Let $(\{x_i^*\}, x^*, y^*)$ denote any limit point of the sequence $\{\{x_i^{t+1}\}, x^{t+1}, y^{t+1}\}$ generated by Algorithm 1, then it is a stationary solution of problem (2.2)
3. If X is a compact set, then the sequence of iterates generated by Algorithm 1 converges to the set of stationary solutions of problem (2.2)

S3: Convergence (cont.)

- To show the first item, notice that

$$
\begin{aligned}
&L(\{x_i^{t+1}\}, x^{t+1}; y^{t+1}) - L(\{x_i^t\}, x^t; y^t) \\
&\le \sum_i \left(\frac{L_i^2}{\rho_i} - \frac{\gamma_i(\rho_i)}{2} \right) \|x_i^{t+1} - x_i^t\|^2 - \frac{\gamma}{2} \|x^{t+1} - x^t\|^2
\end{aligned}
$$

- So the lower boundedness of L implies that

$$
\|x^{t+1} - x^t\| \to 0, \quad \|x_i^{t+1} - x_i^t\| \to 0, \; \forall\; i = 1, \cdots, m. \quad (2.16)
$$

- By the Lemma, we further obtain $\|y_i^{t+1} - y_i^t\| \to 0$ for all $i = 1, 2, ..., m$, which implies that $\|x_i^{t+1} - x^{t+1}\| \to 0$
- The rest of the proof is checking KKT condition, omitted

Discussion

- Comparing with what we seen before, e.g., the DGD method, the assumptions are quite different
    - Convex vs Non-Convex
    - Graph is really simple, fixed, and connected
    - Graph is also special, everyone can talk to the central
    - Non-smooth term only appears in the central node
    - constant parameters ρ_i's
- Is it easy to extend the previous analysis to the general case?
- Not necessarily so, and where is the difficulty?

Numerical Results

- We consider a special case of the consensus problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} x^T B_i x + \lambda \|x\|_1 \\ \text{subject to} \quad & \|x\|_2^2 \leq 1, \end{aligned} \tag{2.17}$$

- Each step closed-form
- $N = 1000$, $m = 10$, $\lambda = 100$. Each $B_k = -\xi\xi^H$, where $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- ρ_k is chosen according to the rule specified in Assumption A2
- We run both the classical and the randomized versions of ADMM, and for the latter case we choose $p_i^t = 0.9$ for all i, t
- We also run the classical ADMM with small stepsizes $\hat{\rho}_i = \rho_i/1000, \ \forall \ i$

Numerical Results

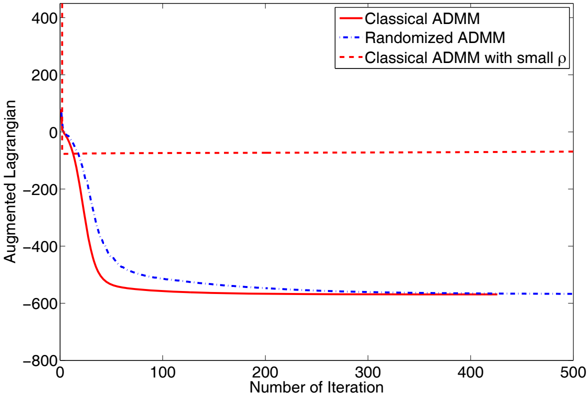


Figure 2.5: The value of $L(x^t; y^t)$ for different algorithms.

Numerical Results

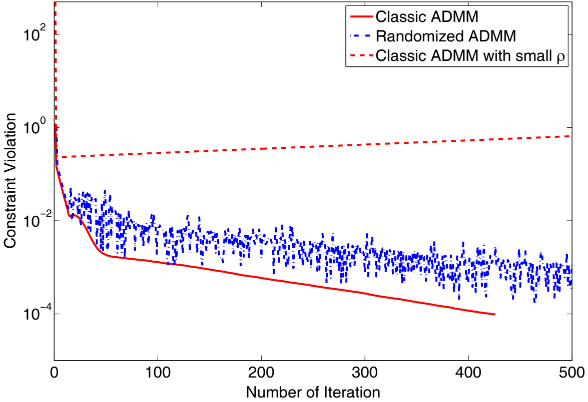


Figure 2.6: The value of $\max_k\{\|x_k^t - x_0^t\|\}$ for different algorithms.

The General Network, and The Primal-Dual Algorithm

More general setting

- Consider a nonconvex decentralized problem

$$\min \quad f(x) := \sum_{i=1}^{m} g_i(x_i) \quad \text{subject to} \quad x_i = x_j, \; (i,j) \in E \tag{3.1}$$

- f_k: smooth, possibly nonconvex function
- No specific assumption on the graph, except that it is fixed, and connected

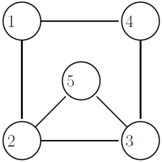
Reformulation

- Introduce local variables $\{x_i\}$, reformulate:

$$
\begin{aligned}
\min_{\{x_i\}} \quad & \sum_{i=1}^{m} f_i(x_i) + h_i(x_i) \\
\text{s.t.} \quad & Ax = 0 \qquad \text{(consensus constraint)}
\end{aligned}
$$

where $A \in \mathbb{R}^{E\times m}$ is the edge-node incidence matrix;
$x := [x_1, \cdots, x_m]^T$

- Recall ,if $e \in \mathcal{E}$ and it connects vertex i and j with $i > j$, then $A_{ev} = 1$ if $v = i$, $A_{ev} = -1$ if $v = j$ and $A_{ev} = 0$ otherwise.



$$
A = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix}.
$$

Linearly Constrained Problem

- So we can take a more generic perspective, by considering the following problem

$$\boxed{\min_{\boldsymbol{x} \in \mathbb{R}^m} \ f(\boldsymbol{x}), \quad \text{subject to} \ A\boldsymbol{x} = b \qquad \text{(Q)}}$$

- Algorithm, analysis and discussion

- Applications and numerical results

The Augmented Lagrangian Algorithm

- We draw elements form AL and Uzawa methods

- The augmented Lagrangian for our problem is given by

$$L_\beta(\boldsymbol{x}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \langle \boldsymbol{\mu}, A\boldsymbol{x} - b \rangle + \frac{\beta}{2}\|A\boldsymbol{x} - b\|^2$$

  where $\boldsymbol{\mu} \in \mathbb{R}^M$ dual variable; $\beta > 0$ penalty parameter

- One primal gradient-type step $+$ one dual gradient-type step

The Proposed Algorithm

- Let $B \in \mathbb{R}^{M\times n}$ be some arbitrary matrix to be defined later
- The Proximal Primal Dual Algorithm [2] is given below

Algorithm 1. The Prox-PDA

At iteration 0, initialize μ^0 and $\boldsymbol{x}^0 \in \mathbb{R}^N$.

At each iteration $r+1$, update variables by:

$$\begin{aligned}\boldsymbol{x}^{r+1} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n}\ & \langle \nabla f(\boldsymbol{x}^r), \boldsymbol{x}-\boldsymbol{x}^r\rangle + \langle \boldsymbol{\mu}^r, A\boldsymbol{x}-b\rangle \\ & + \frac{\beta}{2}\|A\boldsymbol{x}-b\|^2 + \frac{\beta}{2}\|\boldsymbol{x}-\boldsymbol{x}^r\|^2_{B^TB}; \qquad (3.2a)\end{aligned}$$

$$\boldsymbol{\mu}^{r+1} = \boldsymbol{\mu}^r + \beta(A\boldsymbol{x}^{r+1}-b). \qquad (3.2b)$$

[2] M. Hong et al, "Prox-PDA: The Proximal Primal-Dual Algorithm for Fast Distributed Nonconvex Optimization and Learning Over Networks", ICML 2017.

Comments

- The primal iteration has to choose the proximal term

$$\frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{x}^r\|^2_{B^T B}$$

- Choose B appropriately to onensure the following key properties:
  1. The primal problem is strongly convex, hence easily solvable;
  2. The primal problem is decomposable over different variable blocks.

Comments

- Let us illustrate this point
- Consider a network consists of 3 users: $1 \leftrightarrow 2 \leftrightarrow 3$
- Define the graph Laplacian as $L_- = A^T A \in \mathbb{R}^{m \times m}$
- Its (i,i)th diagonal entry is the degree of node i, and its (i,j)th entry is -1 if $e = (i,j) \in \mathcal{E}$, and 0 otherwise.

$$L_- = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}, \qquad L_+ = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

- Define the signless incidence matrix $B := |A|$
- Using this choice of B, we have $B^T B = L_+ \in \mathbb{R}^{m \times m}$, which is the signless graph Laplacian

Comments

- Then objective becomes

$$\sum_{i=1}^{m}\langle \nabla f_i(x_i^r), x_i\rangle + \langle \mu^r, Ax-b\rangle + \frac{\beta}{2}x^T L_- x + \underbrace{\frac{\beta}{2}(x-x^r)^T L_+ (x-x^r)}_{\text{proximal term}}$$

$$= \sum_{i=1}^{m}\langle \nabla f_i(x_i^r), x_i\rangle + \langle \mu^r, Ax-b\rangle + \frac{\beta}{2}x^T (L_- + L_+) x - \beta x^T L_+ x^r$$

$$= \underbrace{\sum_{i=1}^{m}\langle \nabla f_i(x_i^r), x_i\rangle + \langle \mu^r, Ax-b\rangle - \beta x^T L_+ x^r}_{\text{linear in } x} + \beta x^T D x$$

- $D = \text{diag}[d_1, \cdots, d_m] \in \mathbb{R}^{m\times m}$ is the degree matrix
- The problem is separable over the nodes, and strongly convex.

$$\begin{cases} x^{r+1} = \arg\min \langle \nabla f(x^r), x - x^r \rangle + \langle Ax, \mu^r \rangle - \beta x^T L_+ x \\ \qquad + \beta x^T D x \\ \mu^{r+1} = \mu^r + \beta A x^{r+1} \end{cases}$$

$$\equiv \boxed{\text{EXTRA}}$$

$$\nabla f(x^r) + A^T \mu^r - \beta L_+ x^r + 2\beta D x^{r+1} = 0$$

$$\nabla f(x^{r-1}) + A^T \mu^{r-1} - \beta L_+ x^{r-1} + 2\beta D x^r = 0$$

$$0 = \nabla f(x^r) - \nabla f(x^{r-1}) + A^T(\mu^r - \mu^{r-1}) - \beta L_+ (x^r - x^{r-1}) + 2\beta D (x^{r+1} - x^r)$$

$$0 = \nabla f(x^r) - \nabla f(x^{r-1}) + \beta L_- x^r - \beta L_+ (x^r - x^{r-1}) + 2\beta D (x^{r+1} - x^r)$$

$$x^{r+1} = x^r - \frac{1}{2\beta} D^{-1} (\nabla f(x^r) - \nabla f(x^{r-1})) - \frac{1}{2} D^{-1} L_- x^r + \frac{1}{2} D^{-1} L_+ (x^r - x^{r-1})$$

$$= x^r - \frac{1}{2\beta} D^{-1} (\nabla f(x^r) - \nabla f(x^{r-1})) - \frac{1}{2} D^{-1} (L_- - L_+) x^r - \frac{1}{2} D^{-1} L_+ x^{r-1}$$

$$= x^r - \frac{1}{2\beta} D^{-1} (\nabla f^r - \nabla f^{r-1}) + W x^r - \frac{1}{2}(I + W) x^{r-1}$$

$$W = \frac{1}{2} D^{-1} (L^+ - L^-)$$

$$I + W = \underbrace{\frac{1}{2} D^{-1} (L^+ + L^-)}_{2D} + \frac{1}{2} D^{-1} (L^+ - L^-) = D^{-1} L^+$$

Relationship between primal-dual & EXTRA

Compact form of the algorithm

- Can you write down the compact form of the algorithm?
- The relations with the previous algorithms?

Analysis: Assumption

A1. $f(\boldsymbol{x})$ differentiable and has Lipschitz continuous gradient, i.e.,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\mathbf{y})\| \leq L\|\boldsymbol{x} - \mathbf{y}\|, \quad \forall\, \boldsymbol{x}, \mathbf{y} \in \mathbb{R}^N.$$

Further assume that $A^T A + B^T B \succeq I_N$.

A2. There exists a constant $\delta > 0$ such that

$$\exists \underline{f} > -\infty, \quad \text{subject \ to } f(x) + \frac{\delta}{2}\|A\boldsymbol{x} - b\|^2 \geq \underline{f},\ \forall\ \boldsymbol{x} \in \mathbb{R}^N.$$

A3. The constraint $A\boldsymbol{x} = b$ is feasible over $x \in \mathbb{R}^N$.

Functions satisfying the assumption

- **The sigmoid function.** The sigmoid function is given by
$$\mathsf{sigmoid}(x) = \frac{1}{1+e^{-x}} \in [-1,\ 1].$$
- **The** $\arctan$ **function.** $\arctan(x) \in [-1, 1]$ so [A2] is ok. $\arctan'(x) = \frac{1}{x^2+1} \in [0,\ 1]$ so it is bounded, which implies that [A1] is true.
- **The** $\tanh$ **function.** Note that we have
$$\tanh(x) \in [-1, 1], \quad \tanh'(x) = 1 - \tanh(x)^2 \in [0, 1].$$
- **The** logit **function.** The logistic function is related to the $\tanh$ as
$$2\mathsf{logit}(x) = \frac{2e^x}{e^x+1} = 1 + \tanh(x/2).$$
- **The quadratic function** $x^T Q x$**.** Suppose Q is symmetric but not necessarily positive semidefinite, and $x^T Q x$ is strongly convex in the null space of $A^T A$.

The Analysis: Step 1

- Our first step bounds the descent of the augmented Lagrangian
- **Observation**. Dual variable is given as

$$A^T \boldsymbol{\mu}^{r+1} = -\nabla f(\boldsymbol{x}^r) - \beta B^T B(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r)$$

- Change of dual can be bounded by change of primal
- Main idea similar as the previous star-network
- What's the difference?

The Analysis: Step 1

- From the optimality condition of the x problem we have

  $$\nabla f(\boldsymbol{x}^{r+1}) + A^T\boldsymbol{\mu}^r + \beta A^T(A\boldsymbol{x}^{r+1} - b) + \beta B^T B(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r) = 0.$$

  Applying μ update step, we have

  $$A^T\boldsymbol{\mu}^{r+1} = -\nabla f(\boldsymbol{x}^{r+1}) - \beta B^T B(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r). \tag{3.3}$$

- By Assumption [A3], $b \in \mathrm{col}A$; Therefore we must have

  $$\boldsymbol{\mu}^{r+1} - \boldsymbol{\mu}^r = \beta(A\boldsymbol{x}^{r+1} - b) \in \mathrm{col}(A).$$

- This inequality combined with (3.3) implies that

  $$\|\boldsymbol{\mu}^{r+1} - \boldsymbol{\mu}^r\| \le \frac{1}{\sigma_{\min}^{1/2}(A^T A)} \Big\|\nabla f(\boldsymbol{x}^r) - \nabla f(\boldsymbol{x}^{r+1}) - \beta B^T B\left((\boldsymbol{x}^{r+1} - \boldsymbol{x}^r) - (\boldsymbol{x}^r - \boldsymbol{x}^{r-1})\right)\Big\|.$$

The Analysis: Step 1

- Let $\sigma_{\min}(A^T A)$ be the smallest non-zero eigenvalue for $A^T A$

Lemma 3.1

Suppose Assumptions [A1] and [A3] are satisfied. Then:

$$
\begin{aligned}
& L_\beta(\boldsymbol{x}^{r+1}, \boldsymbol{\mu}^{r+1}) - L_\beta(\boldsymbol{x}^r, \boldsymbol{\mu}^r) \\
& \leq -\left(\frac{\beta - L}{2} - \frac{2L^2}{\beta \sigma_{\min}(A^T A)}\right) \|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2 \\
& \quad + \frac{2\beta \|B^T B\|}{\sigma_{\min}(A^T A)} \left\|(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r) - (\boldsymbol{x}^r - \boldsymbol{x}^{r-1})\right\|^2_{B^T B}.
\end{aligned}
$$

For notation simplicity, define

$$
\mathbf{v}^{r+1} := (\boldsymbol{x}^{r+1} - \boldsymbol{x}^r) - (\boldsymbol{x}^r - \boldsymbol{x}^{r-1}). \tag{3.4}
$$

Proof Sketch of Lemma 3.1

- Since $f\boldsymbol{x})$ has Lipschitz continuous gradient, and that $A^TA + B^TB \succeq I$ by Assumption [A1], it is known that if $\beta > L$, then the x-subproblem (3.2a) is strongly convex with modulus $\gamma := \beta - L > 0$;

- That is, we have

$$L_\beta(\boldsymbol{x}, \boldsymbol{\mu}^r) + \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{x}^r\|^2_{B^TB} - (L_\beta(\mathbf{z}, \mu^r) + \frac{\beta}{2}\|\mathbf{z} - \boldsymbol{x}^r\|^2_{B^TB})$$
$$\geq \langle \nabla_x L_\beta(\mathbf{z}, \boldsymbol{\mu}^r) + \beta(B^TB(\mathbf{z} - \boldsymbol{x}^r)), \boldsymbol{x} - \mathbf{z}\rangle + \frac{\gamma}{2}\|\boldsymbol{x} - \mathbf{z}\|^2, \ \forall\ \boldsymbol{x}, \mathbf{z} \in \mathbb{R}^m$$

Proof Sketch of Lemma 3.1

Using this property, we have

$$
\begin{aligned}
& L_\beta(x^{r+1},\mu^{r+1}) - L_\beta(x^r,\mu^r) \\
&= L_\beta(x^{r+1},\mu^{r+1}) - L_\beta(x^{r+1},\mu^r) + L_\beta(x^{r+1},\mu^r) - L_\beta(x^r,\mu^r) \\
&\le L_\beta(x^{r+1},\mu^{r+1}) - L_\beta(x^{r+1},\mu^r) + L_\beta(x^{r+1},\mu^r) + \frac{\beta}{2}\|x^{r+1}-x^r\|^2_{B^TB} - L_\beta(x^r,\mu^r) \\
&\overset{\text{(i)}}{\le} \frac{\|\mu^{r+1}-\mu^r\|^2}{\beta} + \langle \nabla_x L_\beta(x^{r+1},y^r) + \beta(B^TB(x^{r+1}-x^r)), x^{r+1}-x^r\rangle - \frac{\gamma}{2}\|x^{r+1} - x \\
&\overset{\text{(ii)}}{\le} \frac{\|\mu^{r+1}-\mu^r\|^2}{\beta} - \frac{\gamma}{2}\|x^{r+1}-x^r\|^2 \\
&\le \frac{1}{\sigma_{\min}(A^TA)}\left(\frac{2L^2}{\beta}\left\|x^r - x^{r+1}\right\|^2 + 2\beta\left\|B^TB\left((x^{r+1}-x^r)-(x^r-x^{r-1})\right)\right\|^2\right) \\
&\quad - \frac{\gamma}{2}\|x^{r+1}-x^r\|^2 \\
&= -\left(\frac{\beta-L}{2} - \frac{2L^2}{\beta\sigma_{\min}(A^TA)}\right)\|x^{r+1}-x^r\|^2 + \frac{2\beta}{\sigma_{\min}(A^TA)}\left\|B^TBv^{r+1}\right\|^2
\end{aligned}
\tag{3.5}
$$

where in (i) we have used the strong convexity; in (ii) we have used the optimality condition for the x-subproblem (3.2a).

Comments

- Unlike the ADMM for star-network, the rhs cannot be directly made negative
- This suggests that the AL alone does not descend
- Need a new object that is decreasing in the order of

$$\beta \left\|(\boldsymbol{x}^{r+1} - \boldsymbol{x}^{r}) - (\boldsymbol{x}^{r} - \boldsymbol{x}^{r-1})\right\|^2_{B^TB} := \beta \left\|\mathbf{v}^{r+1}\right\|^2_{B^TB}$$

- The change of the sum of the constraint violation $\|A\boldsymbol{x}^{r+1} - b\|^2$ and the proximal term $\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^{r}\|^2_{B^TB}$ has the desired term.

The Analysis: Step 2

Lemma 3.2

Suppose Assumption [A1] is satisfied. Then the following is true

$$\begin{aligned}
&\frac{\beta}{2}\left(\|A\boldsymbol{x}^{r+1}-b\|^2+\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2_{B^TB}\right)\\
&\leq \frac{\beta}{2}\left(\|A\boldsymbol{x}^{r}-b\|^2+\|\boldsymbol{x}^{r}-\boldsymbol{x}^{r-1}\|^2_{B^TB}\right)+L\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\\
&\quad -\frac{\beta}{2}\left(\|\mathbf{v}\|^2_{B^TB}+\|A(\boldsymbol{x}^{r+1}-\boldsymbol{x}^r)\|^2\right).
\end{aligned}$$

- **Observation.** The new object, $\beta/2\left(\|A\boldsymbol{x}^{r+1}-b\|^2+\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2_{B^TB}\right)$, increases in $\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2$ and decreases in $\|\mathbf{v}^{r+1}\|^2_{B^TB}$

The Analysis: Step 2

Lemma 3.2

Suppose Assumption [A1] is satisfied. Then the following is true

$$\begin{aligned}
&\frac{\beta}{2}\left(\|A\boldsymbol{x}^{r+1}-b\|^2+\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2_{B^TB}\right)\\
&\le \frac{\beta}{2}\left(\|A\boldsymbol{x}^{r}-b\|^2+\|\boldsymbol{x}^{r}-\boldsymbol{x}^{r-1}\|^2_{B^TB}\right)+L\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\\
&\quad-\frac{\beta}{2}\left(\|\mathbf{v}\|^2_{B^TB}+\|A(\boldsymbol{x}^{r+1}-\boldsymbol{x}^r)\|^2\right).
\end{aligned}$$

- **Observation.** The new object, $\beta/2\left(\|A\boldsymbol{x}^{r+1}-b\|^2+\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2_{B^TB}\right)$, increases in $\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2$ and decreases in $\|\mathbf{v}^{r+1}\|^2_{B^TB}$
- The change of AL behaves in an opposite manner

The Analysis: Step 2

Lemma 3.2

Suppose Assumption [A1] is satisfied. Then the following is true

$$\begin{aligned}
&\frac{\beta}{2}\left(\|A\boldsymbol{x}^{r+1}-b\|^2+\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2_{B^TB}\right)\\
&\leq \frac{\beta}{2}\left(\|A\boldsymbol{x}^{r}-b\|^2+\|\boldsymbol{x}^{r}-\boldsymbol{x}^{r-1}\|^2_{B^TB}\right)+L\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2\\
&\quad -\frac{\beta}{2}\left(\|\mathbf{v}\|^2_{B^TB}+\|A(\boldsymbol{x}^{r+1}-\boldsymbol{x}^r)\|^2\right).
\end{aligned}$$

- **Observation.** The new object, $\beta/2\left(\|A\boldsymbol{x}^{r+1}-b\|^2+\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2_{B^TB}\right)$, increases in $\|\boldsymbol{x}^{r+1}-\boldsymbol{x}^r\|^2$ and decreases in $\|\mathbf{v}^{r+1}\|^2_{B^TB}$
- The change of AL behaves in an opposite manner
- **Good news.** A conic combination of the two decreases at every iteration.

Derivations

- From the optimality condition of the x-subproblem we have $\forall\ x$

$$\langle \nabla f(\boldsymbol{x}^{r+1}) + A^T\boldsymbol{\mu}^r + \beta A^T(A\boldsymbol{x}^{r+1} - b) + \beta B^T B(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r), \boldsymbol{x}^{r+1} - \boldsymbol{x}\rangle \le 0$$
$$\langle \nabla f(\boldsymbol{x}^{r}) + A^T\boldsymbol{\mu}^{r-1} + \beta A^T(A\boldsymbol{x}^{r} - b) + \beta B^T B(\boldsymbol{x}^{r} - \boldsymbol{x}^{r-1}), \boldsymbol{x}^{r} - \boldsymbol{x}\rangle \le 0.$$

- Plugging $\boldsymbol{x} = \boldsymbol{x}^r$ into the first and $\boldsymbol{x} = \boldsymbol{x}^{r+1}$ into the second, adding, then we obtain

$$\begin{aligned}\langle \nabla f(\boldsymbol{x}^{r+1}) - \nabla f(\boldsymbol{x}^r) + A^T(\boldsymbol{\mu}^{r+1} - \boldsymbol{\mu}^r)\\ + \beta B^T B\left((\boldsymbol{x}^{r+1} - \boldsymbol{x}^r) - (\boldsymbol{x}^r - \boldsymbol{x}^{r-1})\right), \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\rangle \le 0.\end{aligned}$$

- Rearranging, we have

$$\begin{aligned}\langle A^T(\boldsymbol{\mu}^{r+1} - \boldsymbol{\mu}^r), \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\rangle \le -\big\langle \nabla f(\boldsymbol{x}^{r+1}) - \nabla f(\boldsymbol{x}^r)\\ + \beta B^T B\mathbf{v}^{r+1}, \boldsymbol{x}^{r+1} - \boldsymbol{x}^r\big\rangle.\end{aligned} \tag{3.6}$$

Derivations

- Let us bound the lhs and the rhs of (3.6) separately.
- First the lhs of (3.6) can be expressed as

$$
\begin{aligned}
&\langle A^T(\boldsymbol{\mu}^{r+1} - \boldsymbol{\mu}^r), \boldsymbol{x}^{r+1} - \boldsymbol{x}^r \rangle \\
&= \langle \beta A^T (A\boldsymbol{x}^{r+1} - b), \boldsymbol{x}^{r+1} - \boldsymbol{x}^r \rangle \\
&= \langle \beta (A\boldsymbol{x}^{r+1} - b), A\boldsymbol{x}^{r+1} - b - (A\boldsymbol{x}^r - b) \rangle \\
&= \beta \|A\boldsymbol{x}^{r+1} - b\|^2 - \beta \langle A\boldsymbol{x}^{r+1} - b, A\boldsymbol{x}^r - b \rangle \\
&= \frac{\beta}{2} \left(\|A\boldsymbol{x}^{r+1} - b\|^2 - \|A\boldsymbol{x}^r - b\|^2 + \|A(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r)\|^2 \right). \qquad (3.7)
\end{aligned}
$$

- Note: $-ab = -1/2a^2 - 1/2b^2 + 1/2(a-b)^2$

Derivations

- Second we have the following bound for the rhs of (3.6)

$$
\begin{aligned}
&-\langle \nabla f(\boldsymbol{x}^{r+1}) - \nabla f(\boldsymbol{x}^r) + \beta B^T B \mathbf{v}^{r+1}, \boldsymbol{x}^{r+1} - \boldsymbol{x}^r \rangle \\
&\le L\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2 - \beta \langle B^T B \left((\boldsymbol{x}^{r+1} - \boldsymbol{x}^r) - (\boldsymbol{x}^r - \boldsymbol{x}^{r-1}) \right), \boldsymbol{x}^{r+1} - \boldsymbol{x}^r \rangle \\
&= L\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2 + \frac{\beta}{2}\Big(\|\boldsymbol{x}^r - \boldsymbol{x}^{r-1}\|^2_{B^T B} - \|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2_{B^T B} \\
&\qquad - \|(\boldsymbol{x}^r - \boldsymbol{x}^{r-1}) - (\boldsymbol{x}^{r+1} - \boldsymbol{x}^r)\|^2_{B^T B} \Big). \qquad (3.8)
\end{aligned}
$$

- Combining the above two bounds, we have

$$
\begin{aligned}
&\frac{\beta}{2}\left(\|A\boldsymbol{x}^{r+1} - b\|^2 + \|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2_{B^T B} \right) \\
&\le L\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2 + \frac{\beta}{2}\left(\|\boldsymbol{x}^r - \boldsymbol{x}^{r-1}\|^2_{B^T B} + \|A\boldsymbol{x}^r - b\|^2 \right) \\
&\quad - \frac{\beta}{2}\left(\|(\boldsymbol{x}^r - \boldsymbol{x}^{r-1}) - (\boldsymbol{x}^{r+1} - \boldsymbol{x}^r)\|^2_{B^T B} + \|A(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r)\|^2 \right).
\end{aligned}
$$

Step 3. Construction of Potential Functions

- Let us define the potential function for Prox-PDA as

$$P_{c,\beta}^{r+1} = L_\beta(\boldsymbol{x}^{r+1}, \mu^{r+1}) + \frac{c\beta}{2}\left(\|A\boldsymbol{x}^{r+1} - b\|^2 + \|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|_{B^TB}^2\right)$$

where $c > 0$ is some constant to be determined later.

Lemma 3.3

Suppose the assumptions in Lemma 3.2 are satisfied. Then we have

$$\begin{aligned} P_{c,\beta}^{r+1} \le P_{c,\beta}^{r} &- \left(\frac{\beta - L}{2} - \frac{2L^2}{\beta\sigma_{\min}(A^TA)} - cL\right)\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2 \\ &- \left(\frac{c\beta}{2} - \frac{2\beta\|B^TB\|}{\sigma_{\min}(A^TA)}\right)\left\|(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r) - (\boldsymbol{x}^r - \boldsymbol{x}^{r-1})\right\|_{B^TB}^2. \end{aligned}$$

The choice of parameters

- As long as c and β are chosen appropriately, the function $P_{c,\beta}$ decreases at each iteration of Prox-PDA

- The following choices of parameters are sufficient for ensuring descent

$$c \geq \max\left\{\frac{\delta}{L}, \frac{4\|B^T B\|}{\sigma_{\min}(A^T A)}\right\}. \tag{3.9}$$

- The β satisfies

$$\beta > \frac{L}{2}\left(2c+1+\sqrt{(2c+1)^2+\frac{16L^2}{\sigma_{\min}(A^T A)}}\right). \tag{3.10}$$

The Main Result

- Now we are ready to present the main result
- Define $Q(x^{r+1}, \mu^{r+1})$ as the 'stationarity gap'

$$Q(\boldsymbol{x}^{r+1}, \boldsymbol{\mu}^{r}) := \underbrace{\|\nabla_{\boldsymbol{x}} L_\beta(\boldsymbol{x}^{r+1}, \boldsymbol{\mu}^{r})\|^2}_{\text{primal gap}} + \underbrace{\|A\boldsymbol{x}^{r+1} - b\|^2}_{\text{dual gap}}.$$

- $Q(\boldsymbol{x}^{r+1}, \boldsymbol{\mu}^{r}) \to 0$ implies that the limit point $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a KKT point that satisfies the following conditions

$$0 = \nabla f(\boldsymbol{x}^*) + A^T \boldsymbol{\mu}^*, \quad A\boldsymbol{x}^* = b.$$

The Main Result

Theorem 3.4

Suppose Assumption A is satisfied. Further suppose that the conditions on β and c in (3.9) and (3.10) are satisfied. Then

1. **(Eventual Feasibility).** *The constraint is satisfied in the limit:*

$$\lim_{r\to\infty} \boldsymbol{\mu}^{r+1} - \boldsymbol{\mu}^r \to 0, \quad \lim_{r\to\infty} A\boldsymbol{x}^r \to b, \quad \text{and} \quad \lim_{r\to\infty} \boldsymbol{x}^{r+1} - \boldsymbol{x}^r = 0.$$

2. **(Convergence to KKT).** *Every limit point of $\{\boldsymbol{x}^r, \boldsymbol{\mu}^r\}$ converges to a KKT point. Further, $Q(\boldsymbol{x}^{r+1}, \boldsymbol{\mu}^r) \to 0$.*

3. **(Sublinear Convergence Rate).** *For any given $\varphi > 0$, define T to be the first time that the optimality gap reaches below φ, i.e.,*

$$T := \arg\min_r Q(\boldsymbol{x}^{r+1}, \boldsymbol{\mu}^r) \le \varphi.$$

Then there exists a constant $\nu > 0$ such that $\varphi \le \frac{\nu}{T-1}$.

Numerical Results

- We consider the problem of distributed binary classification using nonconvex regularizers in the mini-bach setup
- Each node stores b (batch size) data points, and each component function is given by

$$f_i(x_i) = \frac{1}{Nb}\left[\sum_{j=1}^{b} \log(1+\exp(-y_{ij}x_i^T v_{ij})) + \sum_{k=1}^{M} \frac{\lambda\alpha x_{i,k}^2}{1+\alpha x_{i,k}^2}\right]$$

where $v_{ij} \in \mathbb{R}^M$ and $y_{ij} \in \{1,-1\}$ are the feature vector and the label for the jth date point in ith agent

Numerical Results

- We compare different algorithms with different number of agent in the network (m).
- We measure the optimality gap as well as the constraint violation and the results are respectively reported in Table 1 and Table 2. In the tables Alg1, Alg2, Alg3, Alg4 are denoting Prox-GPDA, Prox-GPDA-IP, DGS, and Push-sum algorithms respectively.

Numerical Results

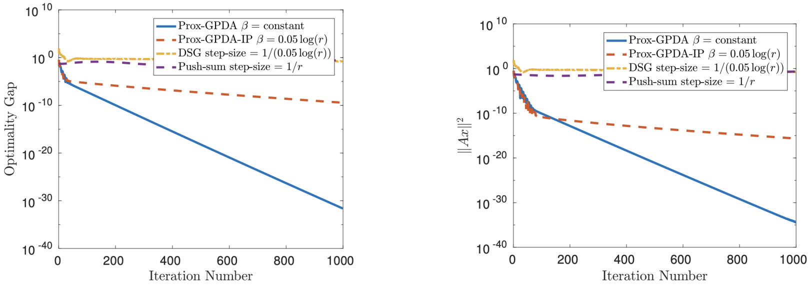


Figure 3.1: The numerical results.

Numerical Results

Table 1: Optimality Gap for different Algorithms

m	**Alg1** (proposed)	**Alg2**	**Alg3**	**Alg4**
10	5.1e-36	2.4e-22	1.34	2.79
20	4.7e-32	5.0e-9	0.04	0.42
30	2.3e-21	5.1e-8	0.008	0.20
40	1.3e-12	2.9e-7	0.007	0.21
50	5.5e-10	4.2e-6	0.005	0.40

Numerical Results

Table 2: Constraint Violation for different Algorithms

m	**Alg1** (proposed)	**Alg2**	**Alg3**	**Alg4**
10	1.3e-36	3.4e-27	0.35	0.65
20	1.2e-34	3.7e-16	0.02	0.40
30	2.3e-24	7.8e-15	0.01	0.18
40	2.2e-16	2.1e-14	0.03	0.20
50	2.2e-14	2.2e-12	0.01	0.12

Summary

- By using ADMM, we can deal with the following problem
  - Non-convex smooth objective function on local agents
  - Constraints/nonsmoothness at the central node
  - Undirected, static and connected graph

- With these settings, the ADMM method, and the primal-dual method, are able to
  1. Converge to desired stationary solutions in the limit
  2. Converge sublinearly to stationary solutions (with $\mathcal{O}(1/T)$ rate)

Discussion

- There are classical works that can deal with non-convex problems, for example, [Tsitsiklis et al 86] [3] [Bianchi - Jakubowicz 13][4]
- But these works do not have rates
- There are also more recent works that can
  - Deal with problems with constraints
  - Deal with more generic graphs/connectivity
  - Deal with stochasticity in the objective
  - ....

[3] J. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," IEEE TAC, 1986

[4] Bianchi, P. and Jakubowicz, "Convergence of a Multi-Agent Projected Stochastic Gradient Algorithm for Non-Convex Optimization", IEEE TAC, 2013

Discussion

- The ADMM/Primal-Dual based methods are useful because
  - They come from a different perspective from the majority of existing optimization problems – the linearly constrained optimization problem
  - Because its optimization roots, it is easier to establish other strongly results (to be discussed later)
  - They have strong connection with the rest of algorithms (such as EXTRA, see the discussion in the previous lecture)

Appendix and Additional Proofs

Proof Sketch of Main Result

- **Step 1.** Bound the size of the gradient of the AL. From the optimality condition of the $\boldsymbol{x}$-problem we have

$$
\begin{aligned}
&\|\nabla_x L_\beta(\boldsymbol{x}^r, \boldsymbol{\mu}^{r-1})\|^2 \\
&= \|\nabla_x L_\beta(\boldsymbol{x}^{r+1}, \boldsymbol{\mu}^r) + \beta B^T B(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r) - \nabla_x L(\boldsymbol{x}^r, \boldsymbol{\mu}^{r-1})\|^2 \\
&= \|\nabla f(\boldsymbol{x}^{r+1}) - \nabla f(\boldsymbol{x}^r) + A^T(\boldsymbol{\mu}^{r+1} - \boldsymbol{\mu}^r) + \beta B^T B(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r)\|^2 \\
&\le 3L^2\|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2 + 3\|\boldsymbol{\mu}^{r+1} - \boldsymbol{\mu}^r\|^2\|A^T A\| + 3\beta^2\|B^T B(\boldsymbol{x}^{r+1} - \boldsymbol{x}^r)\|^2.
\end{aligned}
$$

- By utilizing , we see that there must exist two constants $\xi_1, \xi_2 > 0$ such that the following is true

$$
\begin{aligned}
Q(\boldsymbol{x}^r, \boldsymbol{\mu}^{r-1}) &= \|\nabla_x L_\beta(\boldsymbol{x}^r, \boldsymbol{\mu}^{r-1})\|^2 + \beta\|A\boldsymbol{x}^r - b\|^2 \\
&\le \xi_1 \left\|\boldsymbol{x}^r - \boldsymbol{x}^{r+1}\right\|^2 + \xi_2 \left\|B^T B \mathbf{v}^{r+1}\right\|^2.
\end{aligned}
$$

The last inequalities uses the fact that $\|\boldsymbol{\mu}^{r+1} - \boldsymbol{\mu}^r\|$ can be bounded by $\|B^T B \mathbf{v}^{r+1}\|$, see analysis of Step 1

Proof Sketch

- From the descent estimate we see that there must exist two constants $\nu_1, \nu_2 > 0$ such that

$$\begin{aligned} & P_{c,\beta}(\boldsymbol{x}^{r+1}, \boldsymbol{x}^r, \boldsymbol{\mu}^{r+1}) - P_{c,\beta}(\boldsymbol{x}^r, \boldsymbol{x}^{r-1}, \boldsymbol{\mu}^r) \\ & \leq -\nu_1 \|\boldsymbol{x}^{r+1} - \boldsymbol{x}^r\|^2 - \nu_2 \left\|B^T B \mathbf{v}^{r+1}\right\|^2. \end{aligned}$$

- Matching the above two bounds, we have

$$Q(\boldsymbol{x}^r, \boldsymbol{\mu}^{r-1}) \leq \frac{\min\{\nu_1, \nu_2\}}{\max\{\xi_1, \xi_2\}} \left(P_{c,\beta}(\boldsymbol{x}^r, \boldsymbol{x}^{r-1}, \boldsymbol{\mu}^r) - P_{c,\beta}(\boldsymbol{x}^{r+1}, \boldsymbol{x}^r, \boldsymbol{\mu}^{r+1})\right).$$

- Summing over r, and let T denote the first time that $Q(\boldsymbol{x}^{r+1}, x^r, \boldsymbol{\mu}^{r+1})$ reaches below φ, we obtain

$$\begin{aligned} \varphi & \leq \frac{1}{T-1} \sum_{r=1}^{T} Q(\boldsymbol{x}^r, \boldsymbol{\mu}^{r-1}) \\ & \leq \frac{1}{T-1} \frac{\min\{\nu_1, \nu_2\}}{\max\{\xi_1, \xi_2\}} \left(P_{c,\beta}(\boldsymbol{x}^1, \boldsymbol{x}^0, \boldsymbol{\mu}^1) - P_{c,\beta}(\boldsymbol{x}^{T+1}, \boldsymbol{x}^T, \boldsymbol{\mu}^{T+1})\right) \\ & \leq \frac{1}{T-1} \frac{\min\{\nu_1, \nu_2\}}{\max\{\xi_1, \xi_2\}} \left(P_{c,\beta}(\boldsymbol{x}^1, \boldsymbol{x}^0, \boldsymbol{\mu}^1) - \underline{P}\right) \end{aligned}$$