On the Equivalence of Inexact Proximal ALM and ADMM for a Class of Convex Composite Programming

Defeng Sun

Department of Applied Mathematics

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

DIMACS Workshop on ADMM and
Proximal Splitting Methods in Optimization
June 13, 2018

Joint work with: **Liang Chen** (PolyU), **Xudong Li** (Princeton), and **Kim-Chuan Toh** (NUS)

The multi-block convex composite optimization problem

$$\underbrace{\min_{y\in\mathcal{Y},z\in\mathcal{Z}}}_{w\in\mathcal{W}} \Big\{ \underbrace{p(y_1) + f(y) - \langle b, z\rangle}_{\Phi(w)} \mid \underbrace{\mathcal{F}^* y + \mathcal{G}^* z = c}_{\mathcal{A}^* w = c} \Big\}$$

- $\mathcal{X}$, $\mathcal{Z}$ and $\mathcal{Y}_i$ $(i = 1, \ldots, s)$: finite-dimensional real Hilbert spaces each endowed with $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, $\mathcal{Y} := \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_s$
- $p : \mathcal{Y}_1 \to (-\infty, +\infty]$: a (possibly nonsmooth) closed proper convex function; $f : \mathcal{Y} \to (-\infty, +\infty)$: a continuously differentiable convex function with Lipschitz gradient
- $\mathcal{F}^*$ and $\mathcal{G}^*$: the adjoints of the given linear mappings $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ and $\mathcal{G} : \mathcal{X} \to \mathcal{Z}$
- $b \in \mathcal{Z}$, $c \in \mathcal{X}$: the given data

Too simple? It covers many important classes of convex optimization problems that are best solved in this (dual) form!

A quintessential example

The convex composite quadratic programming (CCQP)

$$\min_{x} \left\{ \psi(x) + \frac{1}{2}\langle x, \mathcal{Q}x\rangle - \langle c, x\rangle \;\middle|\; \mathcal{A}x = b \right\} \tag{1}$$

- $\psi : \mathcal{X} \to (-\infty, +\infty]$: a closed proper convex function
- $\mathcal{Q} : \mathcal{X} \to \mathcal{X}$: a self-adjoint positive semidefinite linear operator

The dual (minimization form):

$$\min_{y_1, y_2, z} \left\{ \psi^*(y_1) + \frac{1}{2}\langle y_2, \mathcal{Q}y_2\rangle - \langle b, z\rangle \;\middle|\; y_1 + \mathcal{Q}y_2 - \mathcal{A}^* z = c \right\} \tag{2}$$

ψ^* is the conjugate of ψ, $y_1 \in \mathcal{X}$, $y_2 \in \mathcal{X}$, $z \in \mathcal{Z}$

- Many problems are subsumed under the convex composite quadratic programming model (1).
- E.g., the important classes of convex quadratic programming (QP), the convex quadratic semidefinite programming (QSDP)...

Convex QSDP

$$\min_{X \in \mathbb{S}^n} \left\{ \frac{1}{2}\langle X, \mathbf{Q}X\rangle - \langle C, X\rangle \;\middle|\; \mathcal{A}_E X = b_E,\ \mathcal{A}_I X \geq b_I,\ X \in \mathbb{S}^n_+ \right\}$$

$\mathbb{S}^n$ is the space of $n \times n$ real symmetric matrices, $\mathbb{S}^n_+$ is the closed convex cone of positive semidefinite matrices in $\mathbb{S}^n$, $\mathbf{Q} : \mathbb{S}^n \to \mathbb{S}^n$ is a positive semidefinite linear operator, $C \in \mathbb{S}^n$ is the given data, and $\mathcal{A}_E$ and $\mathcal{A}_I$ are linear maps from $\mathbb{S}^n$ to certain finite dimensional Euclidean spaces containing b_E and b_I, respectively

- QSDPNAL[1]: a two-phase augmented Lagrangian method in which the first phase is an inexact block sGS decomposition based multi-block proximal ADMM
- The solution generated in the first phase is used as the initial point to warm-start the second phase algorithm

[1] Li, Sun, Toh: QSDPNAL: A two-phase augmented Lagrangian method for convex quadratic semidefinite programming. MPC online (2018)

Penalized and Constrained Regression Models

The penalized and constrained (PAC) regression often arises in high-dimensional generalized linear models with linear equality and inequality constraints, e.g.,

$$\min_{x \in \mathbb{R}^n} \left\{ p(x) + \frac{1}{2\lambda} \|\Phi x - \eta\|^2 \,\middle|\, A_E x = b_E,\ A_I x \geq b_I \right\} \tag{3}$$

- $\Phi \in \mathbb{R}^{m \times n}$, $A_E \in \mathbb{R}^{r_E \times n}$, $A_I \in \mathbb{R}^{r_I \times n}$, $\eta \in \mathbb{R}^m$, $b_E \in \mathbb{R}^{r_E}$ and $b_I \in \mathbb{R}^{r_I}$ are the given data
- p is a proper closed convex regularizer such as $p(x) = \|x\|_1$
- $\lambda > 0$ is a parameter.
- Obviously, the dual of problem (3) is a particular case of CCQP

The augmented Lagrangian function[2]

$$\min_{y\in\mathcal{Y},z\in\mathcal{Z}} \{p(y_1)+f(y)-\langle b,z\rangle \mid \mathcal{F}^*y+\mathcal{G}^*z=c\} \text{ or } \min_{w\in\mathcal{W}} \{\Phi(w) \mid \mathcal{A}^*w=c\}$$

Let $\sigma > 0$ be the penalty parameter. The augmented Lagrangian function:

$$\mathcal{L}_\sigma(y,z;x) := \underbrace{p(y_1)+f(y)-\langle b,z\rangle}_{\Phi(w)} + \underbrace{\langle x, \mathcal{F}^*y+\mathcal{G}^*z-c\rangle}_{\langle x,\mathcal{A}^*w-c\rangle} + \frac{\sigma}{2}\underbrace{\|\mathcal{F}^*y+\mathcal{G}^*z-c\|^2}_{\|\mathcal{A}^*w-c\|^2},$$

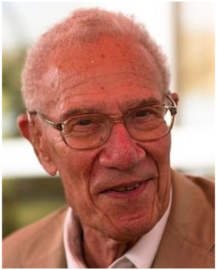
$$\forall w=(y,z)\in\mathcal{W}:=\mathcal{Y}\times\mathcal{Z},\ x\in\mathcal{X}$$

[2] Arrow, K.J., Solow, R.M.: Gradient methods for constrained maxima with weakened assumptions. In: Arrow, K.J., Hurwicz, L., Uzawa, H., (eds.) Studies in Linear and Nonlinear Programming. Stanford University Press, Stanford, pp. 165-176 (1958)

K. Arrow and R. Solow

Kenneth Joseph "Ken" Arrow
(23 August 1921 – 21 February 2017)

John Bates Clark Medal (1957); Nobel Prize in Economics (1972); von Neumann Theory Prize (1986); National Medal of Science (2004); ForMemRS (2006)



Robert Merton Solow
(August 23, 1924 –)

John Bates Clark Medal (1961); Nobel Memorial Prize in Economic Sciences (1987); National Medal of Science (1999); Presidential Medal of Freedom (2014); ForMemRS (2006)

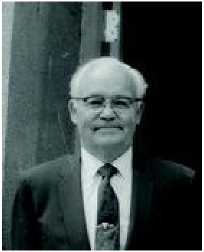
The augmented Lagrangian method[3] (ALM)

$\mathcal{L}_\sigma(y, z; x) = p(y_1) + f(y) - \langle b, z\rangle + \langle x, \mathcal{F}^* y + \mathcal{G}^* z - c\rangle + \frac{\sigma}{2}\|\mathcal{F}^* y + \mathcal{G}^* z - c\|^2$

Starting from $x^0 \in \mathcal{X}$, performs for $k = 0, 1, \ldots$

(1) $\underbrace{(y^{k+1}, z^{k+1})}_{w^{k+1}} \Leftarrow \min_{y,z} \mathcal{L}_\sigma(\underbrace{y, z}_{w}; x^k)$ (approximately)

(2) $x^{k+1} := x^k + \tau\sigma(\mathcal{F}^* y^{k+1} + \mathcal{G}^* z^{k+1} - c)$ with $\tau \in (0, 2)$



Magnus Rudolph Hestenes

(February 13 1906 – May 31 1991)

Michael James David Powell

(29 July 1936 – 19 April 2015)

[3] Also known as the method of multipliers

ALM and variants

- ALM has the desirable asymptotically superlinear convergence (or linearly convergent of an arbitrary order) property.
- While one would really want to $\min_{y,z} \mathcal{L}_\sigma(y, z; x^k)$ without modifying the augmented Lagrangian, it can be expensive due to the coupled quadratic term in y and z.
- In practice, unless the ALM subproblems can be solved efficiently, one would generally want to replace the augmented Lagrangian subproblem with an easier-to-solve surrogate by modifying the augmented Lagrangian function to decouple the minimization with respect to y and z.
- Such a modification is especially desirable during the initial phase of the ALM when the local superlinear convergence phase of ALM has yet to kick in.

ALM to proximal ALM[4] (PALM)

Minimize the augmented Lagrangian function plus a quadratic **proximal term**:

$$w^{k+1} \approx \arg\min_{w} \mathcal{L}_\sigma(w; x^k) + \frac{1}{2}\|w - w^k\|_{\mathcal{D}}^2$$

- $\mathcal{D} = \sigma^{-1}\mathcal{I}$ in the seminal work of Rockafellar (in which inequality constraints are considered). Note that $\mathcal{D} \to 0$ as $\sigma \to \infty$, which is critical for superlinear convergence.
- It is a primal-dual type proximal point algorithm (PPA).

[4] Also known as the proximal method of multipliers

Modification and decomposition

The obvious modification with $\mathcal{D} = \sigma(\lambda^2\mathcal{I} - \mathcal{A}\mathcal{A}^*)$ is generally too drastic and has the undesirable effect of significantly slowing down the convergence of the proximal ALM.

- $\mathcal{D}$ could be positive semidefinite (a kind of PPAs), i.e., the obvious approach:

$$\mathcal{D} = \sigma(\lambda^2\mathcal{I} - \mathcal{A}\mathcal{A}^*) = \sigma(\lambda^2\mathcal{I} - (\mathcal{F};\mathcal{G})(\mathcal{F};\mathcal{G})^*)$$

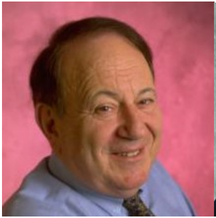
  with λ being the largest singular value of $(\mathcal{F};\mathcal{G})$
- $\mathcal{D}$ can be indefinite (typically used together with the majorization technique)
- **What is an appropriate proximal term** to add so that
  - The PALM subproblem is easier to solve
  - Less drastic than the obvious choice

Decomposition based ADMM

One the other hand, decomposition based approach is available, i.e,

$$y^{k+1} \approx \arg\min_{y}\{\mathcal{L}_\sigma(y, z^k; x^k)\}, \;\; z^{k+1} \approx \arg\min_{z}\{\mathcal{L}_\sigma(y^{k+1}, z; x^k)\}$$

- The **two-block ADMM**
- Allows $\tau \in (0, (1+\sqrt{5})/2)$ if the convergence of the full (primal & dual) sequence is required (Glowinski)
- The case with $\tau = 1$ is a kind of PPA (Gabay + Bertsekas-Eckstein)
- Many variants (proximal/inexact/generalized/parallel etc.)



A part of the result

An equivalent property:

Add an **appropriately designed proximal term** to $\mathcal{L}_\sigma(y, z; x^k)$, we reduce the computation of the modified ALM subproblem to sequentially updating y and z without adding a proximal term, which is **exactly the same** as the two-block ADMM

- A **difference**: one can prove convergence for the step-length τ in the **range** $(0, 2)$ whereas the classic two-block ADMM only admits $(0, (1 + \sqrt{5})/2)$.

For multi-block problems

Turn back to the **multi-block** problem, the subproblem to y can still be difficult due to the coupling of $y_1, \ldots, y_s$

- A successful multi-block ADMM-type algorithm must not only possess **convergence** guarantee but also should numerically **perform** at least as fast as the directly extended ADMM (the Gauss-Seidel iterative fashion) when it does converge.

Algorithmic design

- Majorize the function $f(y)$ at y^k with a quadratic function
- Add an extra proximal term that is derived based on the symmetric Gauss-Seidel (sGS) decomposition theorem to update the sub-blocks in y individually and successively in an sGS fashion
- **The resulting algorithm:**
  A block sGS decomposition based (inexact) majorized multi-block indefinite proximal ADMM with $\tau \in (0, 2)$, which is **equivalent** to an *inexact* majorized proximal ALM

An inexact majorized indefinite proximal ALM

Consider

$$\min_{w \in \mathcal{W}} \Phi(w) := \varphi(w) + h(w) \quad \text{s.t.} \quad \mathcal{A}^* w = c,$$

- The Karush-Kuhn-Tucker (KKT) system:

$$0 \in \partial\varphi(w) + \nabla h(w) + \mathcal{A}x, \qquad \mathcal{A}^* w - c = 0$$

- The gradient of h is Lipschitz continuous, which implies a self-adjoint positive semidefinite linear operator $\widehat{\Sigma}_h : \mathcal{W} \to \mathcal{W}$, such that for any $w, w' \in \mathcal{W}$,

$$h(w) \leq \hat{h}(w, w') := h(w') + \langle \nabla h(w'), w - w' \rangle + \frac{1}{2}\|w - w'\|^2_{\widehat{\Sigma}_h},$$

which is called a majorization of h at w'.

Let $\sigma > 0$. The majorized augmented Lagrangian function is defined, for any $(w, x, w') \in \mathcal{W} \times \mathcal{X} \times \mathcal{W}$, by

$$\widehat{\mathcal{L}}_\sigma(w; (x, w')) := \varphi(w) + \hat{h}(w, w') + \langle \mathcal{A}^* w - c, x \rangle + \frac{\sigma}{2} \|\mathcal{A}^* w - c\|^2.$$

Assumption

The solution set to the KKT system is nonempty and $\mathcal{D} : \mathcal{W} \to \mathcal{W}$ is a given self-adjoint (not necessarily positive semidefinite) linear operator such that

$$\mathcal{D} \succeq -\frac{1}{2}\widehat{\Sigma}_h \quad \textit{and} \quad \frac{1}{2}\widehat{\Sigma}_h + \sigma \mathcal{A}\mathcal{A}^* + \mathcal{D} \succ 0. \tag{4}$$

- **$\mathcal{D}$ is not necessarily to be positive semidefinite!**

Algorithm: an inexact majorized indefinite proximal ALM

Let $\{\varepsilon_k\}$ be a summable sequence of nonnegative numbers. Choose an initial point $(x^0, w^0) \in \mathcal{X} \times \mathcal{W}$. For $k = 0, 1, \ldots,$

1 Compute

$$w^{k+1} \approx \arg\min_{w \in \mathcal{W}} \left\{ \widehat{\mathcal{L}}_\sigma(w; (x^k, w^k)) + \frac{1}{2}\|w - w^k\|_{\mathcal{D}}^2 \right\}$$

such that there exists d_k satisfying $\|d^k\| \leq \varepsilon_k$ and

$$d^k \in \partial_w \widehat{\mathcal{L}}_\sigma(w^{k+1}; (x^k, w^k)) + \mathcal{D}(w^{k+1} - w^k)$$

2 Update $x^{k+1} := x^k + \tau\sigma(\mathcal{A}^* w^{k+1} - c)$ with $\tau \in (0, 2)$

Theorem

The sequence $\{(x^k, w^k)\}$ generated by the above Algorithm converges to a solution to the KKT system.

Multi-block: Majorization and decomposition

The gradient of f is Lipschitz continuous $\Rightarrow$ there exists a self-adjoint linear operator $\widehat{\Sigma}^f : \mathcal{Y} \to \mathcal{Y}$ such that $\widehat{\Sigma}^f \succeq 0$ and for any $y, y' \in \mathcal{Y}$,

$$f(y) \leq \widehat{f}(y, y') := f(y') + \langle \nabla f(y'), y - y' \rangle + \frac{1}{2}\|y - y'\|^2_{\widehat{\Sigma}^f}$$

- Denote for any $y \in \mathcal{Y}$,

$$y_{<i} := (y_1; \ldots; y_{i-1}) \quad \text{and} \quad y_{>i} := (y_{i+1}; \ldots; y_s)$$

- Decompose $\widehat{\Sigma}^f$ as

$$\widehat{\Sigma}^f = \begin{pmatrix} \widehat{\Sigma}^f_{11} & \widehat{\Sigma}^f_{12} & \cdots & \widehat{\Sigma}^f_{1s} \\ (\widehat{\Sigma}^f_{12})^* & \widehat{\Sigma}^f_{22} & \cdots & \widehat{\Sigma}^f_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ (\widehat{\Sigma}^f_{1s})^* & (\widehat{\Sigma}^f_{2s})^* & \cdots & \widehat{\Sigma}^f_{ss} \end{pmatrix}$$

with $\widehat{\Sigma}^f_{ij} : \mathcal{Y}_j \to \mathcal{Y}_i, \ \forall 1 \leq i \leq j \leq s$

Basic assumptions / Majorized augmented Lagrangian

(a) The self-adjoint linear operators $\mathcal{S}_i : \mathcal{Y}_i \to \mathcal{Y}_i, i = 1, \ldots, s$, are chosen such that

$$\frac{1}{2}\widehat{\Sigma}^f_{ii} + \sigma \mathcal{F}_i \mathcal{F}_i^* + \mathcal{S}_i \succ 0 \text{ and } \mathcal{S} := \mathrm{Diag}(\mathcal{S}_1, \ldots, \mathcal{S}_s) \succeq -\frac{1}{2}\widehat{\Sigma}^f$$

(b) The linear operator $\mathcal{G}$ is surjective;

(c) A nonempty solution set to the KKT system:

$$0 \in \begin{pmatrix} \partial p(y_1) \\ 0 \end{pmatrix} + \nabla f(y) + \mathcal{F}x,\ \mathcal{G}x - b = 0,\ \mathcal{F}^* y + \mathcal{G}^* z = c$$

(d) $\{\tilde{\varepsilon}_k\}$ is a summable sequence of nonnegative real numbers

Let $\sigma > 0$. The *majorized* augmented Lagrangian function:

$$\begin{aligned} \widehat{\mathcal{L}}_\sigma(y, z; (x, y')) := \ & p(y_1) + \widehat{f}(y, y') - \langle b, z \rangle \\ & + \langle \mathcal{F}^* y + \mathcal{G}^* z - c, x \rangle + \frac{\sigma}{2} \|\mathcal{F}^* y + \mathcal{G}^* z - c\|^2 \end{aligned}$$

The algorithm sGS-imPADMM

An inexact block sGS based indefinite Proximal ADMM

$(x^0, y^0, z^0) \in \mathcal{X} \times \operatorname{dom} p \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_s \times \mathcal{Z}$. For $k = 0, 1, \ldots,$

1 Compute for $\boxed{i = s, \ldots, 2}$

$$y_i^{k+\frac{1}{2}} \approx \arg\min_{y_i \in \mathcal{Y}_i} \Big\{ \widehat{\mathcal{L}}_\sigma \big(y_{\leq i-1}^k, y_i, y_{\geq i+1}^{k+\frac{1}{2}}, z^k; (x^k, y^k) \big) + \frac{1}{2} \| y_i - y_i^k \|_{\mathcal{S}_i}^2 \Big\}$$

2 Compute for $\boxed{i = 1, \ldots, s}$

$$y_i^{k+1} \approx \arg\min_{y_i \in \mathcal{Y}_i} \Big\{ \widehat{\mathcal{L}}_\sigma \big(y_{\leq i-1}^{k+1}, y_i, y_{\geq i+1}^{k+1/2}, z^k; (x^k, y^k) \big) + \frac{1}{2} \| y_i - y_i^k \|_{\mathcal{S}_i}^2 \Big\}$$

3 Compute

$$z^{k+1} \approx \arg\min_{z \in \mathcal{Z}} \big\{ \widehat{\mathcal{L}}_\sigma (y^{k+1}, z; (x^k, y^k)) \big\}$$

4 Compute $x^{k+1} := x^k + \tau\sigma(\mathcal{F}^* y^{k+1} + \mathcal{G}^* z^{k+1} - c)$, $\boxed{\tau \in (0, 2)}$

Criteria for inexact solutions in sGS-imPADMM

1. For $i = s, \ldots, 2$, the approximate solution $y_i^{k+\frac{1}{2}}$ is chosen such that there exists $\tilde{\delta}_i^k$ satisfying $\|\tilde{\delta}_i^k\| \leq \tilde{\varepsilon}_k$ and

$$\tilde{\delta}_i^k \in \partial_{y_i} \widehat{\mathcal{L}}_\sigma\big(y_{\leq i-1}^k, y_i^{k+\frac{1}{2}}, y_{\geq i+1}^{k+\frac{1}{2}}, z^k; (x^k, y^k)\big) + \mathcal{S}_i(y_i^{k+\frac{1}{2}} - y_i^k)$$

2. For $i = 1, \ldots, s$, the approximate solution y_i^{k+1} is chosen such that there exists δ_i^k satisfying $\|\delta_i^k\| \leq \tilde{\varepsilon}_k$ and

$$\delta_i^k \in \partial_{y_i} \widehat{\mathcal{L}}_\sigma\big(y_{\leq i-1}^{k+1}, y_i^{k+1}, y_{\geq i+1}^{k+1/2}, z^k; (x^k, y^k)\big) + \mathcal{S}_i(y_i^{k+1} - y_i^k)$$

3. The approximate solution z^{k+1} is chosen such that $\|\gamma^k\| \leq \tilde{\varepsilon}_k$ with

$$\begin{aligned} \gamma^k : \quad &= \nabla_z \widehat{\mathcal{L}}_\sigma\big(y^{k+1}, z^{k+1}; (x^k, y^k)\big) \\ &= \mathcal{G}x^k - b + \sigma\mathcal{G}(\mathcal{F}^* y^{k+1} + \mathcal{G}^* z^{k+1} - c) \end{aligned}$$

Comments on the sGS-imPADMM algorithm

- The sGS-imPADMM is a versatile framework, one can implement it in different routines
- We are more interested in the previous iteration scheme:
  - The theoretical improvement
  - The practical merit it features for solving large scale problems (especially when the dominating computational cost is in performing the evaluations associated with the linear mappings $\mathcal{G}$ and $\mathcal{G}^*$)

A particular case in point is the following problem:

$$\min_{x \in \mathcal{X}} \left\{ \psi(x) + \frac{1}{2}\langle x, \mathcal{Q}x \rangle - \langle c, x \rangle \;\middle|\; \mathcal{A}_1 x = b_1,\ \mathcal{A}_2 x \geq b_2 \right\},$$

$\mathcal{Q}$, ψ, and c are as the previous; $\mathcal{A}_1 : \mathcal{X} \to \mathcal{Z}_1$ and $\mathcal{A}_2 : \mathcal{X} \to \mathcal{Z}_2$ are the given linear mappings, and $b = (b_1; b_2) \in \mathcal{Z} := \mathcal{Z}_1 \times \mathcal{Z}_2$ is a given vector.

Details

By introducing a slack variable $x' \in \mathcal{Z}_2$, one gets

$$\min_{x\in\mathcal{X},x'\in\mathcal{Z}_2} \left\{ \psi(x) + \frac{1}{2}\langle x, \mathcal{Q}x\rangle - \langle c, x\rangle \;\middle|\; \begin{pmatrix} \mathcal{A}_1 & 0 \\ \mathcal{A}_2 & \mathcal{I} \end{pmatrix} \begin{pmatrix} x \\ x' \end{pmatrix} = b,\ x' \leq 0 \right\},$$

The corresponding dual problem in the minimization form:

$$\min_{y,y',z} \left\{ p(y) + \frac{1}{2}\langle y', \mathcal{Q}y'\rangle - \langle b, z\rangle \;\middle|\; y + \begin{pmatrix} \mathcal{Q} \\ 0 \end{pmatrix} y' - \begin{pmatrix} \mathcal{A}_1^* & \mathcal{A}_2^* \\ 0 & \mathcal{I} \end{pmatrix} z = \begin{pmatrix} c \\ 0 \end{pmatrix} \right\}$$

with $y := (u, v) \in \mathcal{X} \times \mathcal{Z}_2$, $p(y) = p(u, v) = \psi_1^*(u) + \delta_+(v)$, and δ_+ is the indicator function of the nonnegative orthant in $\mathcal{Z}_2$.

- It is clear that with a large number of inequality constraints, the dimension of z can be much larger than that of y'.
- For such a scenario, the adopted iteration scheme is more preferable since the more difficult subproblem involving z is solved only once in each iteration.

inexact block sGS decomposition

Define $\mathcal{H} := \widehat{\Sigma}^f + \sigma \mathcal{F}\mathcal{F}^* + \mathcal{S} = \mathcal{H}_d + \mathcal{H}_u + \mathcal{H}_u^*$ with
$\mathcal{H}_d := \mathrm{Diag}(\mathcal{H}_{11}, \ldots, \mathcal{H}_{ss})$, $\mathcal{H}_{ii} := \widehat{\Sigma}^f_{ii} + \sigma \mathcal{F}_i \mathcal{F}_i^* + \mathcal{S}_i$ and

$$\mathcal{H}_u := \begin{pmatrix} 0 & \mathcal{H}_{12} & \cdots & \mathcal{H}_{1s} \\ 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \mathcal{H}_{(s-1)s} \\ 0 & 0 & \cdots & 0 \end{pmatrix}, \quad \mathcal{H}_{ij} = \widehat{\Sigma}^f_{ij} + \sigma \mathcal{F}_i \mathcal{F}_j^*$$

For convenience, we denote for each $k \geq 0$, $\tilde{\delta}^1_k := \delta^1_k$, $\tilde{\delta}^k := (\tilde{\delta}^k_1, \tilde{\delta}^2_k \ldots, \tilde{\delta}^k_s)$ and $\delta^k := (\delta^k_1, \ldots, \delta^k_s)$
Define the sequence $\{\Delta^k\} \in \mathcal{Y}$ by

$$\Delta^k := \delta^k + \mathcal{H}_u \mathcal{H}_d^{-1}(\delta^k - \tilde{\delta}^k)$$

Moreover, we can define the linear operator

$$\widehat{\mathcal{H}} := \mathcal{H}_u \mathcal{H}_d^{-1} \mathcal{H}_u^*$$

Result by the block sGS decomposition theorem [5]

The iterate y^{k+1} in Step 2 of sGS-imPADMM is the unique solution to a proximal minimization problem given by

$$y^{k+1} = \arg\min_{y} \Big\{ \underbrace{\widehat{\mathcal{L}}_\sigma(y, z^k; (x^k, y^k)) + \frac{1}{2}\|y - y^k\|^2_{\mathcal{S}+\widehat{\mathcal{H}}}}_{\text{strongly convex}} - \langle \Delta^k, y \rangle \Big\}.$$

Moreover, it holds that

$$\mathcal{H} + \widehat{\mathcal{H}} = (\mathcal{H}_d + \mathcal{H}_u)\mathcal{H}_d^{-1}(\mathcal{H}_d + \mathcal{H}_u^*) \succ 0.$$

- Recall that $\mathcal{H} := \widehat{\Sigma}^f + \sigma\mathcal{F}\mathcal{F}^* + \mathcal{S}$
- Linearly transported error: $\Delta^k = \delta^k + \mathcal{H}_u\mathcal{H}_d^{-1}(\delta^k - \tilde{\delta}^k)$

[5] X.D. Li, D.F. Sun, and K.-C Toh, A block symmetric Gauss-Seidel decomposition theorem for convex composite quadratic programming and its applications, MP online [DOI: 10.1007/s10107-018-1247-7]

The equivalence property

Recall that $\mathcal{W} = \mathcal{Y} \times \mathcal{Z}$. Define $\widehat{\Sigma}_h : \mathcal{W} \to \mathcal{W}$ by

$$\widehat{\Sigma}_h := \begin{pmatrix} \widehat{\Sigma}^f & \\ & 0 \end{pmatrix}$$

For $w = (y; z)$ and $w' = (y'; z')$, denote

$$\widehat{\mathcal{L}}_\sigma(w; (x, w')) := \widehat{\mathcal{L}}_\sigma(y, z; (x, y'))$$

Define the error term

$$\widehat{\Delta}^k := \Delta^k - \mathcal{F}\mathcal{G}^*(\mathcal{G}\mathcal{G}^*)^{-1}(\gamma^{k-1} - \gamma^k - \mathcal{G}(x^{k-1} - x^k)) \in \mathcal{Y}$$

with the convention that

$$\begin{cases} x^{-1} := x^0 - \tau\sigma(\mathcal{F}^* y^0 + \mathcal{G}^* z^0 - c), \\ \gamma^{-1} = -b + \mathcal{G}x^{-1} + \sigma\mathcal{G}(\mathcal{F}^* y^0 + \mathcal{G}^* z^0 - c) \end{cases}$$

The equivalence property

Define the block-diagonal linear operator

$$\mathcal{T} := \begin{pmatrix} \mathcal{S} + \widehat{\mathcal{H}} + \sigma \mathcal{F}\mathcal{G}^*(\mathcal{G}\mathcal{G}^*)^{-1}\mathcal{G}\mathcal{F}^* & \\ & 0 \end{pmatrix} \quad \boxed{\mathcal{W} \to \mathcal{W}}$$

Theorem

Let $\{(x^k, w^k)\}$ with $w^k := (y^k; z^k)$ be the sequence generated by sGS-imPADMM. Then, for any $k \geq 0$, it holds that

(i) *the linear operators $\mathcal{T}$, $\mathcal{A}$ and $\widehat{\Sigma}_h$ satisfy*
$\mathcal{T} \succeq -\frac{1}{2}\widehat{\Sigma}_h$ *and* $\frac{1}{2}\widehat{\Sigma}_h + \sigma\mathcal{A}\mathcal{A}^* + \mathcal{T} \succ 0$;

(ii)

$$w^{k+1} \approx \underset{w \in \mathcal{W}}{\arg\min} \left\{ \widehat{\mathcal{L}}_\sigma\big(w; (x^k, w^k)\big) + \frac{1}{2}\|w - w^k\|^2_{\mathcal{T}} \right\}$$

in the sense that $(\widehat{\Delta}^k; \gamma^k) \in \partial_w \widehat{\mathcal{L}}_\sigma((w^{k+1}; (x^k, w^k)) + \mathcal{T}(w^{k+1} - w^k)$ *and* $\|(\widehat{\Delta}^k, \gamma^k)\| \leq \widehat{\varepsilon}_k$ *with* $\{\widehat{\varepsilon}_k\}$ *being a summable sequence of nonnegative numbers.*

sGS-imPADMM convergence

One can readily get the following convergence theorem

Theorem

The sequence $\{(x^k, y^k, z^k)\}$ *generated by the Algorithm converges to a solution to the KKT system of the problem. Thus,* $\{(y^k, z^k)\}$ *converges to a solution to this problem and* $\{x^k\}$ *converges to a solution of its dual.*

Two-block case

Let $\mathcal{Y} = \mathcal{Y}_1$ and f be vacuous, i.e.,

$$\min\{p(y) - \langle b, z\rangle \,|\, \mathcal{F}^* y + \mathcal{G}^* z = c\} \tag{5}$$

- sGS-imPADMM without proximal terms is reduced to a two-block ADMM
- Assume that $\mathcal{G}$ is surjective and that the KKT system of this problem admits a nonempty solution set K
- This two-block ADMM or its inexact variants with $\tau \in (0, 2)$ (in the order that the y-subproblem is solved before the z-subproblem) converges to K if either $\mathcal{F}$ is surjective or p is strongly convex

Comments on the two-block case

- The assumptions we made for problem (5) are apparently weaker than those in original work of Gabay and Mercier[6], where $\mathcal{F}$ is assumed to be the identity operator and p is assumed to be strongly convex
- In Gabay and Mercier (1976), Theorem 3.1, only the convergence of the primal sequence $\{(y^k, z^k)\}$ is obtained while the dual sequence $\{x^k\}$ is only proven to be bounded
- In Sun *et al.*[7], a similar result to ours has been derived with the requirements that the initial multiplier x^0 satisfies $\mathcal{G}x^0 - b = 0$ and all the subproblems are solved exactly

[6] Gabay, D. and Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)

[7] Sun, D.F., Toh, K.-C. and Yang, L.Q.: A convergent proximal alternating direction method of multipliers for conic programming with 4-block constraints. SIAM J. Optim. **25**(2), 882–915 (2015)

Numerical Experiments

Solving dual linear SDP problems via the two-block ADMM with step-length taking values beyond the standard restriction of $(1+\sqrt{5})/2$. The aim is two-fold.

- As ADMM is among the useful first-order algorithms for solving SDP problems, it is of importance to know to what extent can the numerical efficiency be improved if the equivalence proved in this paper is incorporated.
- As the upper bound of the step-length has been enlarged, it is also important to see whether a step-length that is very close to the upper bound will lead to better or worse numerical performance.

Solving $\min_X \{\langle C, X\rangle \mid \mathcal{A}X = b, X \in \mathbb{S}^n_+\}$,

The dual of the above linear SDP is given by

$$\min_{Y,z} \left\{\delta_{\mathbb{S}^n_+}(Y) - \langle b, z\rangle \mid Y + \mathcal{A}^* z = C\right\},$$

$\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is linear map, $b \in \mathbb{R}^m$ and $C \in \mathbb{S}^n$ are given data.

ADMM has been incorporated in solving dual SDP for a few years

- ADMM with unit step-length was first employed in Povh *et al.* [Comput. 78 (2006)] under the name of boundary point method for solving the dual SDP (Later extended in Malick *et al.* [SIOPT 20 (2009)] with a convergence proof)
- ADMM was used in the software SDPNAL developed by Zhao et al. [SIOPT 20 (2010)] to warm-start a semismooth Newton ALM for dual SDP
- SDPAD by Wen *et al.*[MPC 2 (2010)]: ADMM solver on dual SDP (used SDPNAL template)

ADMM for dual SDP

Let $\sigma > 0$. The augmented Lagrangian function:

$$\mathcal{L}_\sigma(S, z; X) \;=\; \delta_{\mathbb{S}^n_+}(S) - \langle b, z\rangle + \langle X, S + \mathcal{A}^* z - C\rangle + \frac{\sigma}{2}\|S + \mathcal{A}^* z - C\|^2$$

At the k-th step of the two-block ADMM:

$$\begin{cases} S^{k+1} = \Pi_{\mathbb{S}^n_+}(C - \mathcal{A}^* z^k - X^k/\sigma), \\ z^{k+1} = (\mathcal{A}\mathcal{A}^*)^{-1}(\mathcal{A}(C - S^{k+1}) - (\mathcal{A}X^k - b)/\sigma), \\ X^{k+1} = X^k + \tau\sigma(S^{k+1} + \mathcal{A}^* z^{k+1} - C), \end{cases}$$

where $\tau \in (0, 2)$. We emphasize again that this is in contrast to the usual interval of $(0, (1+\sqrt{5})/2)$.

Stopping Criteria: DIMACS[8] rule

Based on relative residuals of priam/dual feasibility and complementarity

We terminate all the tested algorithms if

$$\eta_{\mathsf{SDP}} := \max\{\eta_D, \eta_P, \eta_S\} \leq 10^{-6},$$

where

$$\eta_D = \frac{\|\mathcal{A}^* z + S - C\|}{1+\|C\|}, \eta_P = \frac{\|\mathcal{A}X - b\|}{1+\|b\|}, \eta_S = \max\left\{\frac{\|X - \Pi_{\mathbb{S}^n_+}(X)\|}{1+\|X\|}, \frac{|\langle X, S\rangle|}{1+\|X\|+\|S\|}\right\}$$

with the maximum number of iterations set at 10^6

In addition, we also measure the duality gap:

$$\eta_{\text{gap}} := \frac{\langle C, X\rangle - \langle b, z\rangle}{1 + |\langle C, X\rangle| + |\langle b, z\rangle|}$$

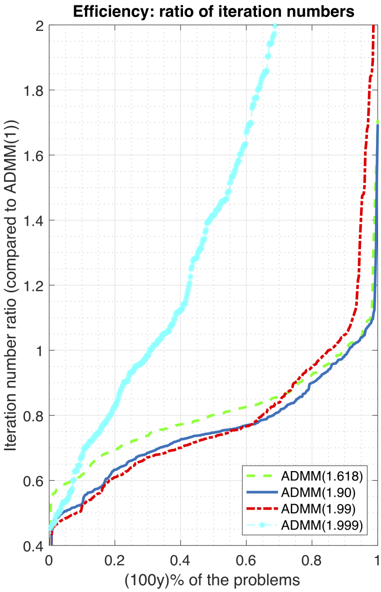
[8] http://dimacs.rutgers.edu/archive/Challenges/Seventh/Instances/error_report.html

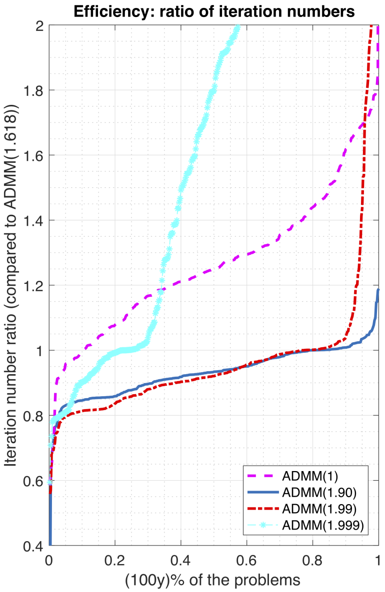
Numerical Experiment: details

- Only consider the cases where $\tau \geq 1$
- We tested five choices of the step-length, i.e., $\tau = 1$, $\tau = 1.618$, $\tau = 1.90$, $\tau = 1.99$ and $\tau = 1.999$
- All these algorithms are tested by running the Matlab package SDPNAL+ (version 1.0)[9]
- We test 6 categories of SDP problems
- In general it is a good idea to use a step-length that is larger than 1, e.g., $\tau = 1.618$, when solving linear SDP problems
- We can even set the step-length to be larger than 1.618, say $\tau = 1.9$, to get better numerical performance

[9] `http://www.math.nus.edu.sg/~mattohkc/SDPNALplus.html`

Numerical result





Conclusions

- For a class of convex composite programming problems, a block sGS decomposition based (inexact) multi-block majorized (proximal) ADMM is equivalent to an inexact proximal ALM.
- An inexact majorized indefinite proximal ALM framework.
- Provide a very general answer to the question on whether the whole sequence generated by the two-block classic ADMM with $\tau \in (0, 2)$, with one linear part, is convergent.
- One can achieve even better numerical performance of the ADMM if the step-length is chosen to be larger than the conventional upper bound of $(1 + \sqrt{5})/2$.
- More insightful theoretical studies on the ADMM-type algorithms are needed for achieving better numerical performance.
- The proximal ALM (with a large proximal term) interpretation of the ADMM may explain why it often converges slow after some iterations.