A block symmetric Gauss-Seidel decomposition theorem and its applications in big data nonsmooth optimization

Defeng Sun

Department of Applied Mathematics

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

The Hong Kong Mathematical Society Annual General Meeting 2018
May 26, 2018

Based on joint works with
Liang Chen (PolyU), Kaifeng Jiang (DBS), Xudong Li (Princeton), Kim-Chuan Toh (NUS) and Liuqin Yang (Grab)

Gauss-Seidel method for solving $\mathrm{Q}\mathbf{x} = \mathbf{b}$

- Update only one element of the variable $\mathbf{x}$ in each iteration.

Input: $\mathrm{Q} \in \Re^{n \times n}, \mathbf{b} \in \Re^n$ and $\mathbf{x}^0 \in \Re^n$

for $k = 0, 1, \ldots$
  for $i = 1, \ldots, n$

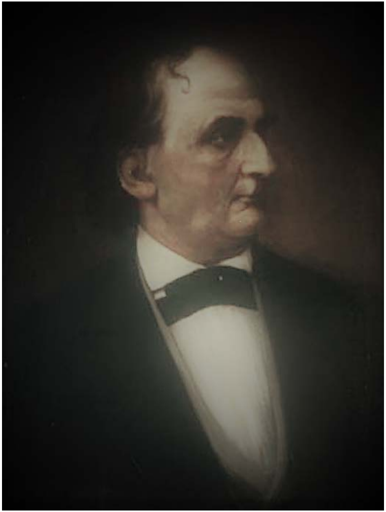
$$\mathbf{x}_i^{k+1} := \mathrm{Q}_{ii}^{-1} \left(\mathbf{b}_i - \sum_{j=1}^{i-1} \mathrm{Q}_{ij} \mathbf{x}_j^{k+1} - \sum_{j=i+1}^{n} \mathrm{Q}_{ij} \mathbf{x}_j^{k} \right)$$

  end for
end for

- converges if Q is diagonally dominant, or symmetric positive definite.

F. Gauss and L. Seidel

Johann Carl Friedrich Gauß
(30 April 1777--23 February 1855)

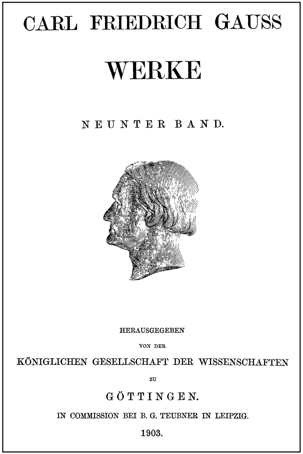


Philipp Ludwig von Seidel
(23 October 1821--13 August 1896)

*Photos from Wikipedia

Gauss-Seidel iteration – resources

Mentioned in a private letter[1] from Gauss to Gerling in 1823.
A publication was not delivered before 1874 by Seidel.



CARL FRIEDRICH GAUSS

WERKE

NEUNTER BAND.

HERAUSGEGEBEN

VON DER

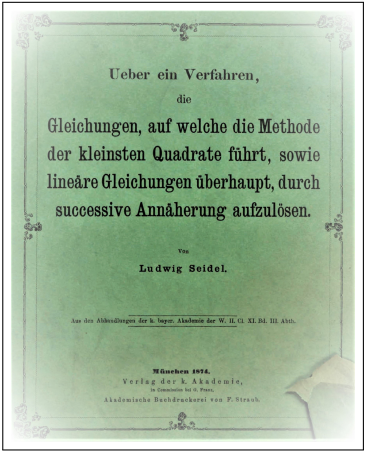
KÖNIGLICHEN GESELLSCHAFT DER WISSENSCHAFTEN

ZU

GÖTTINGEN.

IN COMMISSION BEI B. G. TEUBNER IN LEIPZIG.

1903.



Ueber ein Verfahren,

die

Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineäre Gleichungen überhaupt, durch successive Annäherung aufzulösen.

Von

Ludwig Seidel.

Aus den Abhandlungen der k. bayer. Akademie der W. II. Cl. XI. Bd. III. Abth.

München 1874.

Verlag der k. Akademie,

in Commission bei G. Franz.

Akademische Buchdruckerei von F. Straub.

[1] In Carl Friedrich Gauss Werke 9, Geodäsie, 278-281 (1903). English translation in J.-L. Chabert (Ed.), A History of Algorithms, Springer-Verlag, Berlin, Heidelberg, 297–298 (1999).

[6.]

[Über Stationsausgleichungen.]

Gauss an Gerling. Göttingen, 26. December 1823.

Mein Brief ist zu spät zur Post gekommen und mir zurückgebracht. Ich erbreche ihn daher wieder, um noch die praktische Anweisung zur Elimination beizufügen. Freilich gibt es dabei vielfache kleine Localvortheile, die sich nur ex usu lernen lassen.

......

......

Die Bedingungsgleichungen sind also:

$$
\begin{aligned}
0 &= +6 + 67a - 13b - 28c - 26d \\
0 &= -7558 - 13a + 69b - 50c - 6d \\
0 &= -14604 - 28a - 50b + 156c - 78d \\
0 &= +22156 - 26a - 6b - 78c + 110d;
\end{aligned}
$$

$$\text{Summe} = 0.$$

......

- Gauss considered a 4 dimensional linear equation.
- Starting from $(a, b, c, d) = (0, 0, 0, 0)$, update exactly one variable from $\{a, b, c, d\}$ each time via a certain rule.
- Gauss worked with integers.

Seidel's method

To solve the linear equation

$$\boxed{Ax = b} \quad \text{with} \quad A \in \Re^{m\times n}, b \in \Re^{m},$$

Seidel defined the quadratic function

$$q(x) := \frac{1}{2}\|Ax - b\|_2^2 = \frac{1}{2}\langle x, (A^*A)x\rangle - \langle b, Ax\rangle + \frac{1}{2}\|b\|^2$$

to solve the corresponding normal equation

$$\mathrm{Q}x = A^*b \quad \text{with} \quad \mathrm{Q} := A^*A.$$

- Update only one component of the vector x each step to reduce the value of q.
- The most rational thing (according to Seidel): choose the index that brings the maximum update (decrease) of q.

Gauss-Seidel

The well-known Gauss-Seidel iterative method:

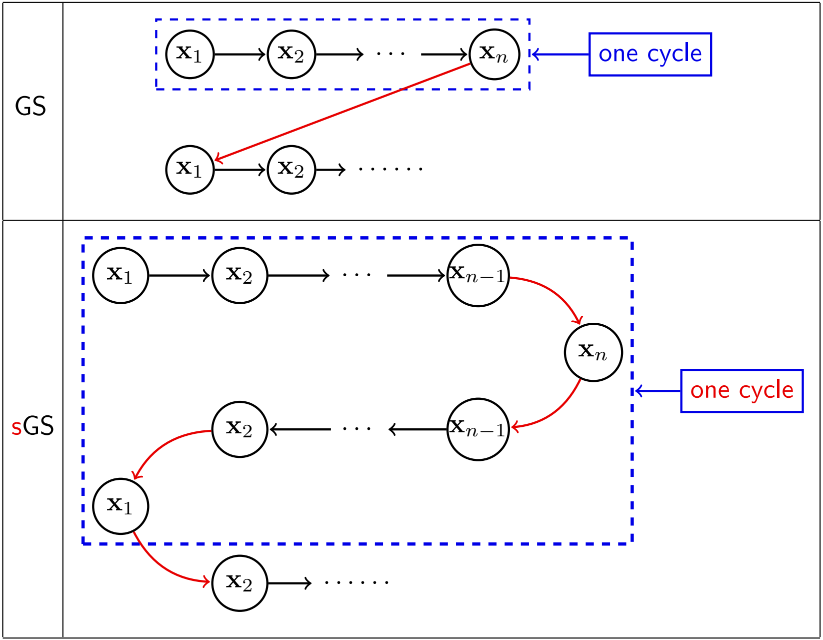
- Forget the "optimal" choice indicated by Gauss and Seidel.
- Changes are carried "cyclically".
- Successively update the elements of x in a fixed order.
- Turn to the first one if the last one is updated.

How about turning to the penultimate one and so on after the last one is updated

- such as the symmetric Gauss-Seidel (sGS) iterative method[2]?

[2] R.E. Bank, T.F. Dupont, and H. Yserentant, "The hierarchical basis multigrid method", Numerische Mathematik 52, 427–458 (1988).

Comparison: GS vs. sGS



A simple optimization model

Let $A \in \Re^{m \times n}$ and $b \in \Re^m$. Let $\mathcal{K} \subseteq \Re^m$ be a closed convex set. Consider the feasibility problem: find $x \in \Re^n$ such that

$$b - Ax \in \mathcal{K},$$

or equivalently, find $x \in \Re^n, z \in \Re^m$ such that

$$z = b - Ax, \quad z \in \mathcal{K}.$$

In the exact spirit as in Seidel's original work, we can consider

$$\min_{(z,x)} \delta_{\mathcal{K}}(z) + \frac{1}{2}\|z + Ax - b\|^2,$$

where $\delta_{\mathcal{K}}(\cdot)$ is the indicator function over $\mathcal{K}$, i.e., $\delta_{\mathcal{K}}(z) = 0$ if $z \in \mathcal{K}$ and $\delta_{\mathcal{K}}(z) = +\infty$ if $z \notin \mathcal{K}$.

- The nonsmooth part $\delta_{\mathcal{K}}(\cdot)$ corresponds to one block of variables!

A general form

Consider the **block** vector
$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s) \in \mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_s$. Given a positive semidefinite linear operator $\mathcal{Q}$ such that

$$\mathcal{Q}\mathbf{x} \equiv \begin{pmatrix} \mathcal{Q}_{11} & \mathcal{Q}_{12} & \cdots & \mathcal{Q}_{1s} \\ \mathcal{Q}_{12}^* & \mathcal{Q}_{22} & \cdots & \mathcal{Q}_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{Q}_{1s}^* & \mathcal{Q}_{2s}^* & \cdots & \mathcal{Q}_{ss} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_s \end{pmatrix}, \quad \mathcal{Q}_{ii} \succ 0.$$

Let $p : \mathcal{X}_1 \to (-\infty, +\infty]$ be a given closed proper convex function. Let the quadratic function

$$q(\mathbf{x}) := \frac{1}{2}\langle \mathbf{x},\, \mathcal{Q}\mathbf{x} \rangle - \langle \mathbf{r},\, \mathbf{x} \rangle.$$

Consider the problem $\boxed{\min_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x_1}) + q(\mathbf{x})}$

- Both block GS and block sGS are applicable.
- block sGS can be used together with the celebrated acceleration technique of Nesterov[3].

[3] Yu. E. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$", Soviet Mathematics Doklady 27(2), 372–376 (1983).

Yurii Nesterov

Yurii Nesterov (January 25, 1956–)

- George Dantzig Prize (2000); John von Neumann Theory Prize (2009); the EURO Gold Medal (2016).
- An accelerated version of the gradient descent method that converges one order faster than the ordinary gradient descent method.

An inexact block sGS iteration

Consider the following block decomposition:

$$\mathcal{U}\mathbf{x} \equiv \begin{pmatrix} 0 & \mathcal{Q}_{12} & \cdots & \mathcal{Q}_{1s} \\ & \ddots & & \vdots \\ & & \ddots & \mathcal{Q}_{(s-1)s} \\ & & & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_s \end{pmatrix}.$$

Then $\mathcal{Q} = \mathcal{U}^* + \mathcal{D} + \mathcal{U}$, where $\mathcal{D}\mathbf{x} = (\mathcal{Q}_{11}\mathbf{x}_1, \ldots, \mathcal{Q}_{ss}\mathbf{x}_s)$.

Let $\hat{\delta} \equiv (\hat{\delta}_1, \ldots, \hat{\delta}_s)$ and $\delta^+ \equiv (\delta_1^+, \ldots, \delta_s^+)$ with $\hat{\delta}_1 = \delta_1^+$ being given error tolerance vectors. Define

$\Delta(\hat{\delta}, \delta^+) := \delta^+ + \mathcal{U}\mathcal{D}^{-1}(\delta^+ - \hat{\delta}),\ \mathcal{T} := \mathcal{U}\mathcal{D}^{-1}\mathcal{U}^*$ (sGS decomp. op.).

Let $\mathbf{y} \in \mathcal{X}$ be given. Define

$$\mathbf{x}^+ := \arg\min_{\mathbf{x}\in\mathcal{X}} \Big\{ p(\mathbf{x}_1) + q(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_{\mathcal{T}}^2 - \langle \Delta(\hat{\delta}, \delta^+), \mathbf{x} \rangle \Big\}. \quad (1)$$

(1) looks complicated, but is much easier to solve!

An inexact block sGS decomposition theorem

Theorem (Li-Sun-Toh)

Given $\mathbf{y}$. *For* $i = s, \ldots, 2$, *define*

$$\begin{aligned} \hat{\mathbf{x}}_i &:= \arg\min_{\mathbf{x}_i} \{ p(\mathbf{y}_1) + q(\mathbf{y}_{\leq i-1}, \mathbf{x}_i, \hat{\mathbf{x}}_{\geq i+1}) - \langle \hat{\delta}_i, \mathbf{x}_i \rangle \} \\ &= \mathcal{Q}_{ii}^{-1} \big(\mathrm{r}_i + \hat{\delta}_i - \textstyle\sum_{j=1}^{i-1} \mathcal{Q}_{ji}^* \mathbf{y}_j - \sum_{j=i+1}^{s} \mathcal{Q}_{ij} \hat{\mathbf{x}}_j \big) \end{aligned}$$

computed in the backward GS cycle. **The optimal solution $\mathbf{x}^+$ in (1) can be obtained exactly via**

$$\begin{aligned} \mathbf{x}_1^+ &= \arg\min_{\mathbf{x}_1} \{ p(\mathbf{x}_1) + q(\mathbf{x}_1, \hat{\mathbf{x}}_{\geq 2}) - \langle \delta_1^+, \mathbf{x}_1 \rangle \}, \\ \mathbf{x}_i^+ &= \arg\min_{\mathbf{x}_i} \{ p(\mathbf{x}_1^+) + q(\mathbf{x}_{\leq i-1}^+, \mathbf{x}_i, \hat{\mathbf{x}}_{\geq i+1}) - \langle \delta_i^+, \mathbf{x}_i \rangle \} \\ &= \mathcal{Q}_{ii}^{-1} (\mathrm{r}_i + \delta_i^+ - \textstyle\sum_{j=1}^{i-1} \mathcal{Q}_{ji}^* \mathbf{x}_j^+ - \sum_{j=i+1}^{s} \mathcal{Q}_{ij} \hat{\mathbf{x}}_j), \quad i \geq 2, \end{aligned}$$

where $\mathbf{x}_i^+$, $i = 1, 2, \ldots, s$, *is computed in the forward GS cycle.*

Reduces to the classical block sGS if both $p(\cdot) \equiv 0$ and $\delta = 0$.
Caution: Such a theorem is not available for GS even if $p(\cdot) \equiv 0$.

An inexact APG (accelerated proximal gradient)

Consider

$$\min\{F(x) := p(\mathbf{x}) + f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$$

with $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \leq L\|\mathbf{x} - \mathbf{x}'\| \quad \forall\, \mathbf{x}, \mathbf{x}' \in \mathcal{X}$.

Algorithm. Input $\mathbf{y}^1 = \mathbf{x}^0 \in \mathrm{dom}(p)$, $t_1 = 1$. Iterate

1. Find an approximate minimizer $\mathbf{x}^k$ to

$$\min_{\mathbf{y}\in\mathcal{X}}\left\{p(\mathbf{y}) + f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), \mathbf{y} - \mathbf{y}^k\rangle + \frac{1}{2}\langle \mathbf{y} - \mathbf{y}^k, \mathcal{H}_k(\mathbf{y} - \mathbf{y}^k)\rangle\right\},$$

   where $\mathcal{H}_k \succ 0$ is a priorily given linear operator.

2. Compute $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$, $\mathbf{y}^{k+1} = \mathbf{x}^k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^k - \mathbf{x}^{k-1})$.

An inexact APG

Consider the following admissible conditions

$$F(\mathbf{x}^k) \ \leq \ p(\mathbf{x}^k) + f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k),\, \mathbf{x}^k - \mathbf{y}^k \rangle + \tfrac{1}{2}\langle \mathbf{x}^k - \mathbf{y}^k,\, \mathcal{H}_k(\mathbf{x}^k - \mathbf{y}^k)\rangle,$$

$$\nabla f(\mathbf{y}^k) + \mathcal{H}_j(\mathbf{x}^k - \mathbf{y}^k) + \gamma^k =: \delta^k \quad \text{with } \|\mathcal{H}_k^{-1/2}\delta^k\| \leq \frac{\epsilon_k}{\sqrt{2}t_k},$$

where $\gamma^k \in \partial p(\mathbf{x}^k)$ = the set of subgradients of p at $\mathbf{x}^k$, $\{\epsilon_k\}$ is a nonnegative summable sequence. Note $t_k \approx k/2$ for k large.

Theorem (Jiang-Sun-Toh)

Suppose that the above conditions hold and $\mathcal{H}_{k-1} \succeq \mathcal{H}_k \succ 0$ for all k. Then

$$0 \leq F(\mathbf{x}^k) - F(\mathbf{x}^*) \leq \frac{4}{(k+1)^2}\Big[\Big(\sqrt{\tau} + \sum_{j=1}^{k}\epsilon_j\Big)^2 + 2\sum_{j=1}^{k}\epsilon_j^2\Big],$$

where $\tau = \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|^2_{\mathcal{H}_1}$.

An inexact APG

Apply the inexact APG to

$$\min\{F(\mathbf{x}) := p(\mathbf{x}_1) + f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}.$$

Since $\nabla f(\cdot)$ is Lipschitz continuous, $\exists$ a symmetric PSD linear operator $\mathcal{Q} : \mathcal{X} \to \mathcal{X}$ such that

$$\mathcal{Q} \ \succeq \ \mathcal{M}, \quad \forall\, \mathcal{M} \in \partial^2 f(\mathbf{x}),\ \forall\, \mathbf{x} \in \mathcal{X}$$

and $\mathcal{Q}_{ii} \succ 0$ for all i.
Given y^k, we have for all $\mathbf{x} \in \mathcal{X}$,

$$f(\mathbf{x}) \ \le \ q_k(\mathbf{x}) := f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k),\, \mathbf{x} - \mathbf{y}^k \rangle + \frac{1}{2}\langle \mathbf{x} - \mathbf{y}^k,\, \mathcal{Q}(\mathbf{x} - \mathbf{y}^k)\rangle.$$

APG subproblem: need to solve a nonsmooth composite QP of the form

$$\min_{\mathbf{x}\in\mathcal{X}}\{p(\mathbf{x}_1) + q_k(\mathbf{x})\}, \quad x = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s),$$

which is not easy to solve!
Idea: add an additional proximal term to make it easier (too easy bad too)!

Elimination of one block via the Danskin theorem

Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s) \in \mathcal{X} := \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_s$ and the corresponding optimization problem

$$
\begin{aligned}
& \min\{p(\mathbf{x}_1) + \varphi(\mathbf{z}) + \phi(\mathbf{z}, \mathbf{x}) \mid \mathbf{z} \in \mathcal{Z},\ \mathbf{x} \in \mathcal{X}\} \\
= \quad & \boxed{\min\{p(\mathbf{x}_1) + f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}},
\end{aligned}
$$

where $p(\cdot)$, $\varphi(\cdot)$ are convex functions (possibly nonsmooth), and

$$
\begin{aligned}
f(\mathbf{x}) &= \min\{\varphi(\mathbf{z}) + \phi(\mathbf{z}, \mathbf{x}) \mid \mathbf{z} \in \mathcal{Z}\}, \\
z(\mathbf{x}) &= \operatorname{argmin}\{\ldots\}.
\end{aligned}
$$

Assume that φ, ϕ satisfy the conditions in the next theorem, then f has Lipschitz continuous gradient $\nabla f(\mathbf{x}) = \nabla_x \phi(\mathbf{z}(\mathbf{x}), \mathbf{x})$.

A Danskin-type theorem

- $\varphi : \mathcal{Z} \to (-\infty, \infty]$ is a closed proper convex function.
- $\phi(\cdot, \cdot) : \mathcal{Z} \times \mathcal{X} \to \Re$ is a convex function.
- $\phi(\mathbf{z}, \cdot) : \Omega \to \Re$ is continuously differentiable on Ω for each $\mathbf{z}$.
- $\nabla_x \phi(\mathbf{z}, \mathbf{x})$ is continuous on $\mathrm{dom}(\varphi) \times \Omega$.

Consider $f : \Omega \to [-\infty, +\infty)$ defined by

$$f(x) = \inf_{\mathbf{z} \in \mathcal{Z}} \{\varphi(\mathbf{z}) + \phi(\mathbf{z}, \mathbf{x})\}, \quad \mathbf{x} \in \Omega.$$

Condition: The minimizer $\mathbf{z}(\mathbf{x})$ is unique for each $\mathbf{x}$ and is bounded on a compact set.

A Danskin-type theorem

Theorem

(i) *If* $\exists$ *an open neighborhood* $\mathcal{N}_{\mathbf{x}}$ *of* $\mathbf{x}$ *such that* $\mathbf{z}(\cdot)$ *is bounded on any compact subset of* $\mathcal{N}_{\mathbf{x}}$, *then the convex function* f *is differentiable on* $\mathcal{N}_{\mathbf{x}}$ *and*

$$\nabla f(\mathbf{x}') = \nabla_{\mathbf{x}} \phi(\mathbf{z}(\mathbf{x}'), \mathbf{x}') \quad \forall\, \mathbf{x}' \in \mathcal{N}_{\mathbf{x}}.$$

(ii) *Suppose that* $\mathbf{z}(\cdot)$ *is bounded on any nonempty compact subset of* $\mathcal{Z}$. *Assume that for any* $\mathbf{z} \in \mathrm{dom}(\varphi)$, $\nabla_{\mathbf{x}} \phi(\mathbf{z}, \cdot)$ *is Lipschitz continuous on* $\mathcal{Z}$ *and* $\exists\, \Sigma \succeq 0$ *such that for all* $\mathbf{x} \in \mathcal{X}$ *and* $\mathbf{z} \in \mathrm{dom}(\varphi)$,

$$\Sigma \succeq \mathcal{H} \quad \forall\, \mathcal{H} \in \partial^2_{\mathbf{xx}} \phi(\mathbf{z}, \mathbf{x}).$$

Then, $\nabla f(\cdot)$ *is Lipschitz continuous on* $\mathcal{X}$ *with the Lipschitz constant* $||\Sigma||_2$ *(the spectral norm of* Σ*) and for any* $\mathbf{x} \in \mathcal{X}$,

$$\Sigma \succeq \mathcal{G} \quad \forall\, \mathcal{G} \in \partial^2 f(\mathbf{x}),$$

where $\partial^2 f(\mathbf{x})$ *denotes the generalized Hessian of* f *at* $\mathbf{x}$.

$$\min\{p(\mathbf{x}_1) + \varphi(\mathbf{z}) + \phi(\mathbf{z}, \mathbf{x}) \mid \mathbf{z} \in \mathcal{Z},\ \mathbf{x} \in \mathcal{X}\}$$

Algorithm 2. Input $\mathbf{y}^1 = \mathbf{x}^0 \in \mathrm{dom}(p) \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_s$, $t_1 = 1$. Let $\{\epsilon_k\}$ be a nonnegative summable sequence. Iterate

1. Suppose δ_i^k, $\hat{\delta}_i^k \in \mathcal{X}_i$, $i = 1, \ldots, s$, with $\hat{\delta}_1^k = \delta_1^k$, are error vectors such that

$$\max\{\|\delta^k\|, \|\hat{\delta}^k\|\} \le \epsilon_k/(\sqrt{2}t_k),$$

$$\mathbf{z}^k = \arg\min_{\mathbf{z}} \left\{\varphi(\mathbf{z}) + \phi(\mathbf{z}, \mathbf{y}^k)\right\}, \quad \text{(elimination via Danskin)}$$

$$\mathbf{x}^k = \arg\min_{\mathbf{x}} \left\{p(\mathbf{x}_1) + q_k(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{y}^k\|_{\mathcal{T}}^2 - \langle \Delta(\hat{\delta}^k, \delta^k), \mathbf{x}\rangle\right\}.$$

(inexact sGS)

2. Compute $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$, $\mathbf{y}^{k+1} = \mathbf{x}^k + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^k - \mathbf{x}^{k-1})$.

An inexact accelerated block coordinate gradient descent

Theorem

Let $\mathcal{H} = \mathcal{Q} + \mathcal{T}$ *and* $\beta = 2\|\mathcal{D}^{-1/2}\| + \|\mathcal{H}^{-1/2}\|$. *The sequence* $\{(\mathbf{z}^k, \mathbf{x}^k)\}$ *generated by Algorithm 2 satisfies*

$$0 \leq F(\mathbf{x}^k) - F(\mathbf{x}^*) \leq \frac{4}{(k+1)^2}\Big[\Big(\sqrt{\tau} + \beta\sum_{j=1}^{k}\epsilon_j\Big)^2 + 2\beta^2\sum_{j=1}^{k}\epsilon_j^2\Big],$$

where $\tau = \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_{\mathcal{H}}^2$.

Least squares semidefinite programming (LSSDP)

Given fixed G, g, consider the LSSDP

$$\begin{aligned} \min \quad & \boldsymbol{F(Z, v, S, y_E, y_I)} := [\delta^*_{\mathcal{P}}(-Z) + \delta^*_{\mathcal{K}}(-v)] + \delta_{\mathcal{S}^n_+}(S) \\ & -\langle b_E,\, y_E\rangle + \frac{1}{2}\|Z + S + \mathcal{A}^*_E y_E + \mathcal{A}^*_I y_I + G\|^2 + \frac{1}{2}\|v - y_I + g\|^2, \end{aligned}$$

where for a given closed convex set $\mathcal{C}$, $\delta^*_{\mathcal{C}}(\cdot)$ is the conjugate function of $\delta_{\mathcal{C}}(\cdot)$ defined by

$$\delta^*_{\mathcal{C}}(\cdot) = \sup_{W\in\mathcal{C}} \langle \cdot,\, W\rangle,$$

$\mathcal{S}^n_+$ is the cone of n by n symmetric positive semidefinite matrices, and $\mathcal{P}$ is a polyhedral set.

Existing first-order methods for LSSDP

- Block coordinate descent (BCD) type method [Luo,Tseng,...] with iteration complexity of $O(1/k)$.
- Accelerated proximal gradient (APG) method [Nesterov, Beck-Teboulle] with iteration complexity of $O(1/k^2)$.
- Accelerated randomized BCD-type method [Beck, Nesterov, Richtarik,...] with iteration complexity of $O(1/k^2)$.

Inexact ABCD for LSSDP: version 1

Step 1. Suppose $\delta_E^k,\ \hat{\delta}_E^k \in \Re^{m_E}$, $\delta_I^k,\ \hat{\delta}_I^k \in \Re^{m_I}$ satisfy

$$\max\{\|\delta_E^k\|, \|\delta_I^k\|, \|\hat{\delta}_E^k\|, \|\hat{\delta}_I^k\|\} \le \frac{\epsilon_k}{\sqrt{2}t_k}.$$

$$\begin{aligned}
(Z^k, v^k) &= \arg\min_{Z,v}\big\{F(Z, v, \widetilde{S}^k, \widetilde{y}_E^k, \widetilde{y}_I^k)\big\}, && \text{(Projection onto } \mathcal{P},\ \mathcal{K}\text{)} \\
\hat{y}_E^k &= \arg\min_{y_E}\big\{F(Z^k, v^k, \widetilde{S}^k, y_E, \widetilde{y}_I^k) - \langle \hat{\delta}_E^k,\ y_E\rangle\big\}, && \text{(Chol. or CG)} \\
\hat{y}_I^k &= \arg\min_{y_I}\big\{F(Z^k, v^k, \widetilde{S}^k, \hat{y}_E^k, y_I) - \langle \hat{\delta}_I^k,\ y_I\rangle\big\}, && \text{(Chol. or CG)} \\
S^k &= \arg\min_{S}\big\{F(Z^k, v^k, S, \hat{y}_E^k, \hat{y}_I^k)\big\}, && \text{(Projection onto } \mathbb{S}_+^n\text{)} \\
y_I^k &= \arg\min_{y_I}\big\{F(Z^k, v^k, S^k, \hat{y}_E^k, y_I) - \langle \delta_I^k,\ y_I\rangle\big\}, && \text{(Chol. or CG)} \\
y_E^k &= \arg\min_{y_E}\big\{F(Z^k, v^k, S^k, y_E, y_I^k) - \langle \delta_E^k,\ y_E\rangle\big\}. && \text{(Chol. or CG)}
\end{aligned}$$

Step 2. Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ and $\tau_k = \frac{t_k - 1}{t_{k+1}}$. Compute

$$(\widetilde{S}^{k+1}, \widetilde{y}_E^{k+1}, \widetilde{y}_I^{k+1}) = (1+\tau_k)(S^k, y_E^k, y_I^k) - \tau_k(S^{k-1}, y_E^{k-1}, y_I^{k-1}).$$

Inexact ABCD for LSSDP: version 2

We can also treat (S, y_E, y_I) as a single block and use a semismooth Newton-CG (SNCG) algorithm introduced in [Zhao-Sun-Toh, SIAM J. Optim. 20(4), 1737-1765 (2010)] to solve it inexactly. Choose $\tau = 10^{-6}$.

Step 1. Suppose $\delta_E^k \in \Re^{m_E}$, $\delta_I^k \in \Re^{m_I}$ are error vectors such that

$$\max\{\|\delta_E^k\|, \|\delta_I^k\|\} \leq \frac{\epsilon_k}{\sqrt{2}t_k}.$$

Compute

$$(Z^k, v^k) = \arg\min_{Z,v} \left\{F(Z, v, \widetilde{S}^k, \widetilde{y}_E^k, \widetilde{y}_I^k)\right\}, \quad \text{(Projection onto } \mathcal{P}, \mathcal{K}\text{)}$$

$$(S^k, y_E^k, y_I^k) = \arg\min_{S, y_E, y_I} \left\{ \begin{array}{c} F(Z^k, v^k, S, y_E, y_I) + \frac{\tau}{2}\|y_E - \widetilde{y}_E^k\|^2 \\ -\langle \delta_E^k,\, y_E\rangle - \langle \delta_I^k,\, y_I\rangle \end{array} \right\}.$$

(SNCG)

Step 2. Set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$, $\tau_k = \frac{t_k - 1}{t_{k+1}}$. Compute

$$(\widetilde{S}^{k+1}, \widetilde{y}_E^{k+1}, \widetilde{y}_I^{k+1}) = (1 + \tau_k)(S^k, y_E^k, y_I^k) - \tau_k(S^{k-1}, y_E^{k-1}, y_I^{k-1}).$$

Numerical experiments

- We compare the performance of ABCD against BCD, APG and eARBCG (an enhanced accelerated randomized block coordinate gradient method) for solving LSSDP.
- We test the algorithms on LSSDP problem by taking $G = -C$, $g = 0$ for the data arising from various classes of semidefinite programming (SDP).

Numerical results

Stop the algorithms after 25,000 iterations, or

$$\eta = \max\{\eta_1, \eta_2, \eta_3\} < 10^{-6},$$

where $\eta_1 = \frac{\|b_E - \mathcal{A}_E X\|}{1+\|b_E\|}$, $\eta_2 = \frac{\|X-Y\|}{1+\|X\|}$, $\eta_3 = \frac{\|s-\mathcal{A}_I X\|}{1+\|s\|}$,

$X = \Pi_{\mathbb{S}^n_+}(\mathcal{A}_E^* y_E + \mathcal{A}_I^* y_I + Z + G)$, $Y = \Pi_{\mathcal{P}}(\mathcal{A}_E^* y_E + \mathcal{A}_I^* y_I + S + G)$, $s = \Pi_{\mathcal{K}}(g - y_I)$.

problem set (No.) \ solver	ABCD	APG	eARBCG	BCD
θ_+ (64)	64	64	64	11
FAP (7)	7	7	7	7
QAP (95)	95	95	24	0
BIQ (165)	165	165	165	65
RCP (120)	120	120	120	108
exBIQ (165)	165	141	165	10
Total (616)	616	592	545	201

Performance profiles

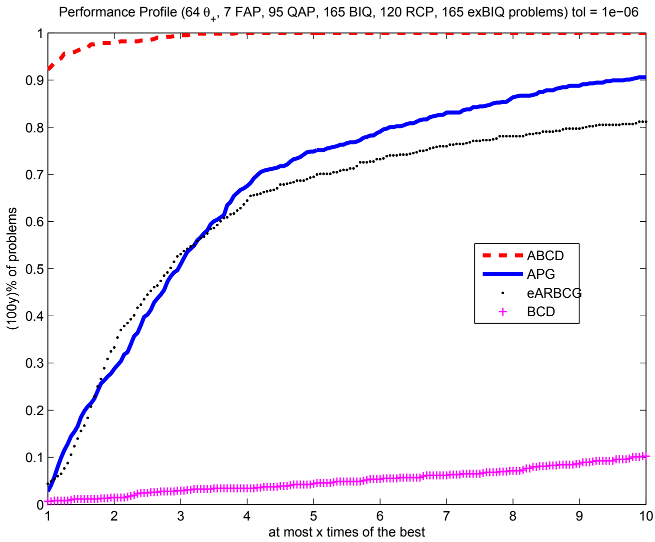


Figure: Performance profiles of ABCD, APG, eARBCG and BCD on $[1, 10]$

Higher accuracy results for ABCD

Number of problems which are solved to the accuracy of 10^{-6}, 10^{-7}, 10^{-8} by the ABCD method.

problem set (No.)	10^{-6}	10^{-7}	10^{-8}
θ_+ (64)	64	58	52
FAP (7)	7	7	7
QAP (95)	95	95	95
BIQ (165)	165	165	165
RCP (120)	120	120	118
exBIQ (165)	165	165	165
Total (616)	616	610	602

Tolerance profiles of the ABCD

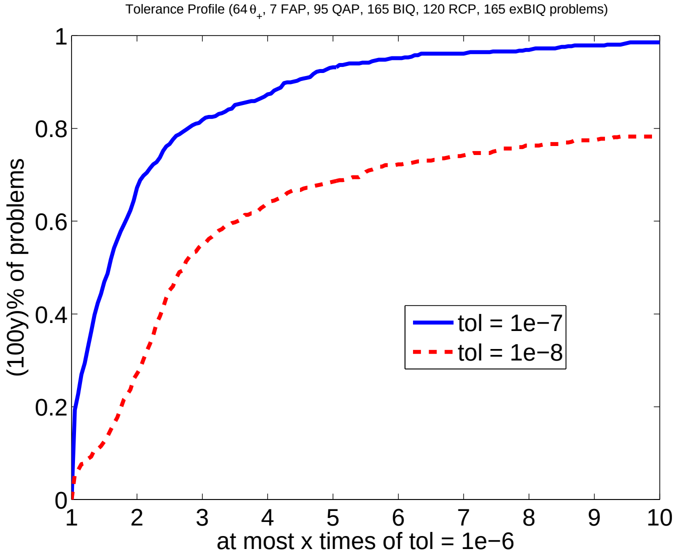


Figure: Tolerance profiles of ABCD on $[1, 10]$

Consider the convex optimization model:

$$
\begin{array}{ll}
\min & \theta(y_1) + f(y_1, y_2, \ldots, y_s) \\
\text{s.t.} & \mathcal{A}_1^* y_1 + \mathcal{A}_2^* y_2 + \cdots + \mathcal{A}_s^* y_s = c\,.
\end{array} \tag{2}
$$

Linear mappings: $\mathcal{A}_i,\ i = 1, \ldots, s,\ \mathcal{A}^* y = \sum_{i=1}^{s} \mathcal{A}_i^* y_i,\ y := (y_1, \ldots, y_s)$. Closed proper convex function $\theta : \mathcal{Y}_1 \to (-\infty, +\infty]$ and convex quadratic function $f(y) = \frac{1}{2}\langle y,\, \mathcal{Q}y\rangle - \langle b,\, y\rangle$. Then, (2) can be written compactly as

$$
\min\{\theta(y_1) + f(y) \mid \mathcal{A}^* y = c\}.
$$

Given $\sigma > 0$, the augmented Lagrangian function of the CCQP is

$$
\mathcal{L}_\sigma(y; x) = \theta(y_1) + \underbrace{f(y) + \langle x,\, \mathcal{A}^* y - c\rangle + \frac{\sigma}{2}\|\mathcal{A}^* y - c\|^2}_{\text{quadratic}}\,.
$$

The proximal augmented Lagrangian method (pALM) for the CCQP:

Given (y^0, x^0) in the domain and $\tau \in (0,2)$. For $k = 0, 1, \ldots$

Step 1. $y^{k+1} \approx \arg\min \mathcal{L}_\sigma(y; x^k) + \frac{1}{2}\|y - y^k\|^2_{\mathcal{T}}$

$$= \arg\min_y \left\{\theta(y_1) + f(y) + \langle x^k, \mathcal{A}^* y - c\rangle + \frac{\sigma}{2}\|\mathcal{A}^* y - c\|^2 + \frac{1}{2}\|y - y^k\|^2_{\mathcal{T}}\right\}.$$

Step 2. $x^{k+1} = x^k + \tau\sigma(\mathcal{A}^* y^{k+1} - c)$.

- $\mathcal{T}$ is the block sGS decomposition operator, which does not need to be formulated explicitly. Note that $\mathcal{T} \succeq 0$ but $\mathcal{T} \not\succ 0$. So it is not a classical pALM.
- y^{k+1} is obtained via the inexact block sGS procedure [s blocks in total].
- In practice, the dual step-length τ is often chosen in $[1.618, 1.95]$.

Consider the convex composite quadratic programming

$$\min_{x\in\mathcal{X}}\Big\{\psi(x)+\frac{1}{2}\langle x,\mathcal{Q}x\rangle-\langle c,x\rangle \;\Big|\; \mathcal{A}_E x=b_E,\ \mathcal{A}_I x\geq b_I\Big\}. \tag{3}$$

- $\psi:\mathcal{X}\to(-\infty,+\infty]$ is a closed proper convex function.
- $\mathcal{Q}:\mathcal{X}\to\mathcal{X}$ is a self-adjoint positive semidefinite linear operator.
- $\mathcal{A}_E:\mathcal{X}\to\mathcal{Z}_1$ and $\mathcal{A}_I:\mathcal{X}\to\mathcal{Z}_2$ are the given linear mappings.
- $b=(b_E;b_I)\in\mathcal{Z}:=\mathcal{Z}_1\times\mathcal{Z}_2$ is a given vector.
- $c\in\mathcal{X}$, $b\in\mathcal{Z}$ are the given data.

Let $\mathcal{I}$ be the identity operator in $\mathcal{Z}_2$. By introducing a slack variable $x'\in\mathcal{Z}_2$, we can reformulate the above problem equivalently as

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

whose dual is an instance of the CCQP (in the next page).

The dual of the above problem [or equivalently problem (3)] is

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

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

- Nonsmoothness only exists in one block of variables, i.e., the y-block.
- Block sGS + pALM work perfectly [both y' and z can be decomposed into many blocks].
- Convex quadratic programming (QP), Convex quadratic semidefinite programming (QSDP), ...

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.,

$$\min_{x \in \mathbb{R}^n} \left\{ p(x) + \frac{1}{2\lambda} \|\Phi x - \eta\|^2 \,\middle|\, A_E x = b_E,\ A_I x \geq b_I \right\}. \tag{4}$$

- $\Phi \in \mathbb{R}^{m \times n}$, $A_E \in \mathbb{R}^{r_E \times n}$, $A_I \in \mathbb{R}^{r_I \times n}$, $\eta \in \mathbb{R}^m$, $b_E \in \mathbb{R}^{r_E}$ 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 (4), which is a special case of problem (3), is a particular case of CCQP.

Extensions

- There are many applications that can be "solved" via block sGS + pALM if the solution accuracy is not a big concern.
- More extensions can be done. For example, for the doubly non-negative SDP problems or the rank-correction models, for the dual forms (more efficient in general), one needs to deal with TWO nonsmooth blocks plus many smooth blocks. Then, again, one can use the sGS decomposition theorem + proximal ADMM (pADMM) instead of pALM to handle these situations [not often encountered in optimization applications].
- As one can see, we can also deal with problems whose objective functions involving non-quadratic smooth functions via majorizations.
- To make the algorithms even faster, we often introduce indefinite proximal terms with guaranteed convergence.
- Here, for big sparse optimization problems, the more critical second order sparsity (SOS) is not touched yet ...

References

- X.Y. Zhao, D.F. Sun, K.C. Toh, A Newton-CG augmented Lagrangian method for semidefinite programming, SIAM J. Optimization 20 (2010) 1737-1765
- K.F. Jiang, D.F. Sun, K.C. Toh, An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP, SIAM J. Optimization 22 (2012) 1042–1064
- D.F. Sun, K.C. Toh and L.Q. Yang, A convergent 3-block semi-proximal ADMM for conic programming with 4-type constraints, SIAM J. Optimization 25 (2015) 882–915
- X.D. Li, D.F. Sun, K.C. Toh, A Schur complement based semiproximal ADMM for convex quadratic conic programming and extensions, Mathematical Programming 155 (2016) 333–373
- D.F. Sun, K.C. Toh, L.Q. Yang, An efficient inexact ABCD method for least squares semidefinite programming, SIAM J. Optimization 26 (2016) 1072–1100

- L. Chen, D.F. Sun, K.C. Toh, An efficient inexact symmetric Gauss-Seidel based majorized ADMM for high-dimensional convex composite conic programming, Mathematical Programming 161 (2017) 237–270
- X.D. Li, D.F. Sun, K.C. Toh, A block sGS decomposition theorem for convex composite quadratic programming and its applications, Mathematical Programming (2018) 1–24 DOI:10.1007/s10107-018-1247-7
- X.D. Li, D.F. Sun, K.C. Toh, QSDPNAL: A two-phase augmented Lagrangian method for convex quadratic SDP, Mathematical Programming Computation (2018) 1–41 DOI:10.1007/s12532-018-0137-6
- L. Chen, X.D. Li, D.F. Sun, K.C. Toh, On the equivalence of inexact proximal ALM and ADMM for a class of convex composite programming, arXiv:1803.10803 (2018)

Thank you for your attention!