Matrix optimization:
recent progress on algorithm foundation

Chao Ding
Institute of Applied Mathematics
Academy of Mathematics and Systems Science
Chinese Academy of Sciences

ISMP2018, Bordeaux, France
July 6, 2018
Acknowledgements

Based on the joint work with **Ying Cui** at **USC** and **Xinyuan Zhao** at **BISEC, Beijing**.


Matrix optimization problem

Augmented Lagrangian method
  The convex case
  The non-convex case

Two perturbation results
  The calmness of $X$
  The strong regularity of the general MOP
Matrix optimization problem
Matrix Optimization Problem (MOP)

MOP:
\[
\min_{x \in X} \quad f(x) + \theta(g(x)) \\
\text{s.t.} \quad h(x) \in Q
\]
Matrix Optimization Problem (MOP)

MOP:

$$\min_{x \in \mathbb{X}} \ f(x) + \theta(g(x))$$

s.t. \quad h(x) \in \mathcal{Q}

- \mathbb{X}: a finite dimensional Euclidean space
**Matrix Optimization Problem (MOP)**

\[
\begin{align*}
\text{MOP:} & \quad \min_{x \in X} \ f(x) + \theta(g(x)) \\
& \quad \text{s.t.} \quad h(x) \in Q
\end{align*}
\]

- **X**: a finite dimensional Euclidean space
- **f**: a smooth function (e.g., the data-fitting term)
Matrix Optimization Problem (MOP)

MOP:
\[
\min_{x \in X} f(x) + \theta(g(x)) \\
\text{s.t. } h(x) \in Q
\]

- $X$: a finite dimensional Euclidean space
- $f$: a smooth function (e.g., the data-fitting term)
- $\theta$: a function defined on the finite dimensional matrix space (symmetric/Hermitian or non-symmetric/non-Hermitian)
Matrix Optimization Problem (MOP)

MOP:

\[
\min_{x \in X} f(x) + \theta(g(x)) \\
\text{s.t. } h(x) \in Q
\]

- **X**: a finite dimensional Euclidean space
- **f**: a smooth function (e.g., the data-fitting term)
- **θ**: a function defined on the finite dimensional matrix space (symmetric/Hermitian or non-symmetric/non-Hermitian)
  - the regularization term (structured features, e.g., lower rank), e.g., \(\delta_{S^n_{\pm}}(\cdot), \| \cdot \|_{\ast}, \| \cdot \|_{(k)}, \text{Schatten norm.}\)
- **g** and **h**: smooth functions
Matrix Optimization Problem (MOP)

**MOP:**

\[
\min_{x \in X} f(x) + \theta(g(x)) \\
\text{s.t. } h(x) \in Q
\]

- \(X\): a finite dimensional Euclidean space
- \(f\): a smooth function (e.g., the data-fitting term)
- \(\theta\): a function defined on the finite dimensional matrix space (symmetric/Hermitian or non-symmetric/non-Hermitian)
  the regularization term (structured features, e.g., lower rank), e.g., \(\delta_{S_n^+}(\cdot), \| \cdot \|_{\ast}, \| \cdot \|_{(k)}\), Schatten norm.
- \(g\) and \(h\): smooth functions
- \(Q\): a convex polyhedral set
The regularization term and has the following composite structure:

\[ \theta(X) = \phi(\sigma(X)) \]
The regularization term and has the following composite structure:

\[ \theta(X) = \phi(\sigma(X)) \]

- \( \phi: \mathbb{R}^m \to (-\infty, +\infty] \) is an (absolutely) symmetric function
- \( \sigma(X) = (\sigma_1(X), \ldots, \sigma_m(X)) \) with \( \sigma_1(X) \geq \ldots \geq \sigma_m(X) \): spectral (singular values/eigenvalues) of \( X \)
The **regularization term** and has the following **composite structure**:

\[ \theta(X) = \phi(\sigma(X)) \]

- \( \phi : \mathbb{R}^m \rightarrow (-\infty, +\infty] \) is an *(absolutely) symmetric* function
- \( \sigma(X) = (\sigma_1(X), \ldots, \sigma_m(X)) \) with \( \sigma_1(X) \geq \ldots \geq \sigma_m(X) \): **spectral** (singular values/eigenvalues) of \( X \)

e.g., \( \delta_{S^+}; \delta_{\text{EDM}}; \| \cdot \|_*; \) the spectral norm \( \| \cdot \|_2; \) Ky Fan’s \( k \)-norm \( \| \cdot \|_{(k)} \)
The regularization term and has the following composite structure:

\[ \theta(X) = \phi(\sigma(X)) \]

- \( \phi : \mathbb{R}^m \rightarrow (-\infty, +\infty] \) is an (absolutely) symmetric function
- \( \sigma(X) = (\sigma_1(X), \ldots, \sigma_m(X)) \) with \( \sigma_1(X) \geq \ldots \geq \sigma_m(X) \): spectral (singular values/eigenvalues) of \( X \)

E.g., \( \delta_{S^n} \); \( \delta_{\text{EDM}} \); \( \| \cdot \|_* \); the spectral norm \( \| \cdot \|_2 \); Ky Fan’s \( k \)-norm \( \| \cdot \|_{(k)} \)

Matrix completion:

\[
\min \left\{ \| P_\Omega(X) - P_\Omega(M) \|^2_F + \lambda \| X \|_* \mid \mathcal{A}(X) = b, \; X \in \mathcal{P} \right\}
\]

\( P_\Omega(\cdot) \): the observation operator
More applications

• SDP
• Fastest mixing Markov chain problem (fast load balancing of paralleled systems)
• Fastest distributed linear averaging problem
• The reduced rank approximations of transition matrices
• The low rank approximations of doubly stochastic matrices
• Low-rank approximation of matrices with linear structures
• ......
More applications

- SDP
- Fastest mixing Markov chain problem (fast load balancing of paralleled systems)
- Fastest distributed linear averaging problem
- The reduced rank approximations of transition matrices
- The low rank approximations of doubly stochastic matrices
- Low-rank approximation of matrices with linear structures
- ......
- **Unsupervised learning**
MOPs in unsupervised learning

The fastest mixing Markov chain problem:¹

- find the edge transition probabilities that give the fastest mixing Markov chain
- equivalent to minimize the second largest eigenvalue modulus of $P$

$$\min \|P\|_{(2)}$$

s.t. $P \geq 0$, $P1 = 1$, $P = P^T$

$P_{ij} = 0$, $(i, j) \notin \mathcal{E}$

---

Numerical results

Table 1: The performance of the augmented Lagrangian method (ALM) and the ADMM for solving the FMMC problems. The computational time is in the format of “hh:mm:ss”.

<table>
<thead>
<tr>
<th>problem</th>
<th>$d; n$</th>
<th>iteration</th>
<th>$\eta$</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>cage</td>
<td>2562 ; 366</td>
<td>6;6;200</td>
<td>0.0-7</td>
<td>05</td>
</tr>
<tr>
<td>G3</td>
<td>19176 ; 800</td>
<td>32;57;88</td>
<td>3.0-7</td>
<td>1:37</td>
</tr>
<tr>
<td>G6</td>
<td>9665 ; 800</td>
<td>30;44;145</td>
<td>8.5-7</td>
<td>1:09</td>
</tr>
<tr>
<td>G15</td>
<td>4661 ; 800</td>
<td>31;51;200</td>
<td>3.4-8</td>
<td>1:05</td>
</tr>
<tr>
<td>G46</td>
<td>9990 ; 1000</td>
<td>30;44;134</td>
<td>5.6-7</td>
<td>1:34</td>
</tr>
<tr>
<td>G54</td>
<td>5916 ; 1000</td>
<td>22;62;200</td>
<td>7.7-7</td>
<td>2:23</td>
</tr>
<tr>
<td>G43</td>
<td>9990 ; 1000</td>
<td>24;96;90</td>
<td>2.9-7</td>
<td>2:37</td>
</tr>
<tr>
<td>delaunayn10</td>
<td>3056 ; 1024</td>
<td>61;359;200</td>
<td>6.8-9</td>
<td>10:25</td>
</tr>
<tr>
<td>G22</td>
<td>19990 ; 2000</td>
<td>31;46;56</td>
<td>2.3-8</td>
<td>5:57</td>
</tr>
<tr>
<td>G24</td>
<td>19990 ; 2000</td>
<td>41;296;200</td>
<td>2.7-7</td>
<td>53:57</td>
</tr>
<tr>
<td>G26</td>
<td>19990 ; 2000</td>
<td>29;87;200</td>
<td>1.4-7</td>
<td>16:15</td>
</tr>
<tr>
<td>minnesota</td>
<td>3303 ; 2642</td>
<td>25;24;123</td>
<td>0.0-10</td>
<td>6:27</td>
</tr>
<tr>
<td>G48</td>
<td>6000 ; 3000</td>
<td>40;79;200</td>
<td>9.2-7</td>
<td>19:39</td>
</tr>
<tr>
<td>G49</td>
<td>6000 ; 3000</td>
<td>25;38;200</td>
<td>6.6-7</td>
<td>10:58</td>
</tr>
<tr>
<td>G50</td>
<td>6000 ; 3000</td>
<td>26;42;74</td>
<td>4.8-8</td>
<td>9:36</td>
</tr>
<tr>
<td>USpowerGrid</td>
<td>6594 ; 4941</td>
<td>27;120;200</td>
<td>1.4-7</td>
<td>3:11:22</td>
</tr>
</tbody>
</table>
Why ALM?

A quick answer: it just works!

ALM has the asymptotically superlinear convergence / linearly convergent of an arbitrary order.
A quick answer: it just works!
A quick answer: it just works!

**ALM** has the *asymptotically superlinear* convergence / *linearly convergent* of an *arbitrary* order.
Canonically perturbed MOPs with parameters \((a, b, c)\):

\[
\begin{align*}
\text{min} & \quad f(x) - \langle a, x \rangle + \theta(g(x) + b) \\
\text{s.t.} & \quad h(x) + c \in Q
\end{align*}
\]
**Perturbed MOPs**

Canonically perturbed MOPs with parameters \((a,b,c)\):

\[
\begin{align*}
\min \ & f(x) - \langle a, x \rangle + \theta(g(x) + b) \\
\text{s.t.} \ & h(x) + c \in Q
\end{align*}
\]

Under suitable CQs, the **Karush-Kuhn-Tucker (KKT)** optimality condition for perturbed problem takes the following form:

\[
\begin{align*}
\begin{cases}
\mathbf{a} &= \nabla f(x) + g'(x)^* y + h'(x)^* z \\
y &\in \partial \theta(g(x) + b) &\iff \text{GE} : \ (a, b, c) \in T_L(x, y, z) \\
z &\in N_K(h(x) + c)
\end{cases}
\end{align*}
\]
Perturbed MOPs

Canonically perturbed MOPs with parameters \((a,b,c)\):

\[
\begin{align*}
\min & \quad f(x) - \langle a, x \rangle + \theta(g(x) + b) \\
\text{s.t.} & \quad h(x) + c \in Q
\end{align*}
\]

Under suitable CQs, the Karush-Kuhn-Tucker (KKT) optimality condition for perturbed problem takes the following form:

\[
\begin{cases}
    a = \nabla f(x) + g'(x)^* y + h'(x)^* z \\
    y \in \partial \theta(g(x) + b) \\
    z \in N_K(h(x) + c)
\end{cases}
\implies \text{GE}: (a, b, c) \in T_L(x, y, z)
\]

\(X(a, b, c)\): the set of all locally optimal solutions
Canonicaally perturbed MOPs with parameters $(a,b,c)$:

$$\min f(x) - \langle a, x \rangle + \theta(g(x) + b)$$
$$\text{s.t. } h(x) + c \in Q$$

Under suitable CQs, the Karush-Kuhn-Tucker (KKT) optimality condition for perturbed problem takes the following form:

$$\begin{cases}
    a = \nabla f(x) + g'(x)^* y + h'(x)^* z \\
y \in \partial \theta(g(x) + b) \\
z \in N_K(h(x) + c)
\end{cases} \iff \text{GE} : (a, b, c) \in T_L(x, y, z)$$

$X(a, b, c)$: the set of all locally optimal solutions

$S_{\text{KKT}}(a, b, c)$: the set of all solutions $(x, y, z)$ to the KKT system
Let $\Psi$ be a set-valued mapping with $(\overline{X}, \overline{Y}) \in \text{gph} \, \Psi$. Lipschitz-like properties:

- **Calmness**: there exist neighborhoods $U$ of $\overline{X}$, $V$ of $\overline{Y}$ and constant $\kappa > 0$ such that $\Psi(\overline{X}) \cap V \subseteq \Psi(\overline{X}) + \kappa \| \overline{X} - \overline{X} \|_B \forall \overline{X} \in U$.

- **Robust isolated calmness**: there exist neighborhoods $U$ of $\overline{X}$, $V$ of $\overline{Y}$ and constant $\kappa > 0$ such that $\emptyset \neq \Psi(\overline{X}) \cap V \subseteq \{ \overline{Y} \} + \kappa \| \overline{X} - \overline{X} \|_B \forall \overline{X} \in U$.

- **Strong regularity**: $\text{SKKT}(a, b, c)$ is locally Lipschitz continuous.
Let $\Psi$ be a \textit{set-valued mapping} with $(\overline{X}, \overline{Y}) \in \text{gph} \, \Psi$. \textit{Lipschitz-like} properties:

- \textbf{Calmness:}
  
  there exist neighborhoods $U$ of $\overline{X}$, $V$ of $\overline{Y}$ and constant $\kappa > 0$ such that
  
  $$\Psi(X) \cap V \subseteq \Psi(\overline{X}) + \kappa \|X - \overline{X}\|_{\mathcal{B}} \quad \forall X \in U.$$
Let $\Psi$ be a set-valued mapping with $(\overline{X}, \overline{Y}) \in \text{gph} \, \Psi$. **Lipschitz-like** properties:

- **Calmness:**
  there exist neighborhoods $U$ of $\overline{X}$, $V$ of $\overline{Y}$ and constant $\kappa > 0$ such that
  \[
  \Psi(X) \cap V \subseteq \Psi(\overline{X}) + \kappa \|X - \overline{X}\|_B \quad \forall \, X \in U.
  \]

- **Robust isolated calmness:**
  there exist neighborhoods $U$ of $\overline{X}$, $V$ of $\overline{Y}$ and constant $\kappa > 0$ such that
  \[
  \emptyset \neq \Psi(X) \cap V \subseteq \{\overline{Y}\} + \kappa \|X - \overline{X}\|_B \quad \forall \, X \in U.
  \]
Let $\Psi$ be a set-valued mapping with $(\overline{X}, \overline{Y}) \in \text{gph} \, \Psi$. Lipschitz-like properties:

- **Calmness:**
  there exist neighborhoods $U$ of $\overline{X}$, $V$ of $\overline{Y}$ and constant $\kappa > 0$ such that
  \[ \Psi(X) \cap V \subseteq \Psi(\overline{X}) + \kappa \|X - \overline{X}\|_B \ \forall X \in U. \]

- **Robust isolated calmness:**
  there exist neighborhoods $U$ of $\overline{X}$, $V$ of $\overline{Y}$ and constant $\kappa > 0$ such that
  \[ \emptyset \neq \Psi(X) \cap V \subseteq \{\overline{Y}\} + \kappa \|X - \overline{X}\|_B \ \forall X \in U. \]

**Strong regularity:**

$S_{KKT}(a, b, c)$ is locally Lipschitz continuous.
Why it matters?

- **Theory**: perturbation analysis of OPs
- **Algorithm**: e.g., convergence rate of augmented Lagrangian method
Why it matters?

- **Theory**: perturbation analysis of OPs
Why it matters?

- **Theory**: perturbation analysis of OPs
- **Algorithm**: e.g., convergence rate of augmented Lagrangian method
Augmented Lagrangian method
Augmented Lagrangian method (ALM)

Initiated by Hestenes (1969); Powell (1969)

- Convex case: PPA
  - global convergence
  - local (super)linear convergence

Rockafellar (1976a), (1976b); Pennanen (2002); ...
Augmented Lagrangian method (ALM)

Initiated by Hestenes (1969); Powell (1969)

- Convex case: PPA
  - global convergence
  - local (super)linear convergence
  Rockafellar (1976a), (1976b); Pennanen (2002); ...

- Non-convex case: for NLP, under suitable conditions, local convergence at a linear rate
  Powell (1972); Rockafellar (1973a), (1973b); Tretyakov (1973), Bertsekas (1976); Conn et al. (1991), Contesse-Back (1993); Ito and Kunisch (1990); ...
Augmented Lagrangian method (ALM)

Initiated by Hestenes (1969); Powell (1969)

- Convex case: PPA
  - global convergence
  - local (super)linear convergence

  Rockafellar (1976a), (1976b); Pennanen (2002); ...

- Non-convex case: for NLP, under suitable conditions, local convergence at a linear rate

  Powell (1972); Rockafellar (1973a), (1973b); Tretyakov (1973), Bertsekas (1976); Conn et al. (1991), Contesse-Back (1993); Ito and Kunisch (1990); ...

A survey paper:

Augmented Lagrangian method

The convex case

The non-convex case
Consider the following quadratic MOP and its dual:

\[
(P) \quad \min \quad \frac{1}{2} \| \mathcal{F}X - d \|^2 + \langle C, X \rangle + \theta(X) \\
\text{s.t.} \quad \mathcal{A}X = b, \; X \in \mathcal{P}
\]

\[
(D) \quad \min \quad \frac{1}{2} \| w \|^2 - \langle w, d \rangle - \langle b, y \rangle + \theta^*(-S) + \delta^*_P(-Z) \\
\text{s.t.} \quad \mathcal{F}^*w + \mathcal{A}^*y + S + Z = C
\]
The convex case

Consider the following quadratic MOP and its dual:

\[(P) \quad \min_{X} \quad \frac{1}{2} \| \mathcal{F}X - d \|^2 + \langle C, X \rangle + \theta(X) \]
\[\text{s.t.} \quad AX = b, \ X \in \mathcal{P} \]

\[(D) \quad \min_{w} \quad \frac{1}{2} \| w \|^2 - \langle w, d \rangle - \langle b, y \rangle + \theta^*(-S) + \delta^*_P(-Z) \]
\[\text{s.t.} \quad \mathcal{F}^*w + \mathcal{A}^*y + S + Z = C \]

The **Lagrange function** of (D):

\[L(w, y, S, Z, X) := \frac{1}{2} \| w \|^2 - \langle b, y \rangle - \langle d, w \rangle + \theta^*(-S) + \delta^*_P(-Z) \]
\[+ \langle X, \mathcal{F}^*w + \mathcal{A}^*y + S + Z - C \rangle, \]
The augmented Lagrangian function associated with (D) is:

\[ L_\sigma (w, y, S, Z, X) := L(w, y, S, Z, X) + \frac{\sigma}{2} \| F^* w + A^* y + S + Z - C \|^2. \]

ALM for solving (D):

Given a sequence of scalars \( \sigma_k \uparrow \sigma_\infty \leq \infty \) and a starting point \( X^0 \in \mathbb{R}^{m \times n} \), the \((k + 1)\)-th iteration of the inexact ALM:

\[
\begin{cases}
    z^{k+1} := (w^{k+1}, y^{k+1}, S^{k+1}, Z^{k+1}) \approx \text{arg min } \zeta_k(z) := L_{\sigma_k}(w, y, S, Z, X^k) \\
    X^{k+1} = X^k + \sigma_k(F^* w^{k+1} + A^* y^{k+1} + S^{k+1} + Z^{k+1} - C)
\end{cases}
\]
The **augmented Lagrangian function** associated with (D)

\[ L_\sigma(w, y, S, Z, X) := L(w, y, S, Z, X) + \frac{\sigma}{2}\|\mathcal{F}^* w + \mathcal{A}^* y + S + Z - C\|^2. \]

**ALM for solving (D):**

Given a sequence of scalars \( \sigma_k \uparrow \sigma_\infty \leq \infty \) and a starting point \( X^0 \in \mathbb{R}^{m \times n} \), the \((k+1)\)-th iteration of the inexact ALM:

\[
\begin{align*}
  z^{k+1} := (w^{k+1}, y^{k+1}, S^{k+1}, Z^{k+1}) &\approx \arg \min \zeta_k(z) := L_{\sigma_k}(w, y, S, Z, X^k) \\
  X^{k+1} &= X^k + \sigma_k(\mathcal{F}^* w^{k+1} + \mathcal{A}^* y^{k+1} + S^{k+1} + Z^{k+1} - C)
\end{align*}
\]

Terminate the subproblem for solving \( z^{k+1} \) by one of following criteria:

(A) \( \zeta_k(z^{k+1}) - \inf \zeta_k(z) \leq \varepsilon_k^2/2\sigma_k, \varepsilon_k \geq 0, \sum_{k=0}^\infty \varepsilon_k < \infty \)

(B) \( \zeta_k(z^{k+1}) - \inf \zeta_k(z) \leq (\eta_k^2/2\sigma_k)\|X^{k+1} - X^k\|^2, \eta_k \geq 0, \sum_{k=0}^\infty \eta_k < \infty \)
ALM & PPA

ALM for (D) $\iff$ PPA (proximal point algorithm) for (P)
ALM & PPA

ALM for (D) $\iff$ PPA (proximal point algorithm) for (P)

The essential objective function of (P)

$$F(X) := \inf \left\{ L(w, y, S, Z, X) \mid (w, y, S, Z) \in \mathbb{R}^d \times \mathbb{R}^e \times \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \right\}$$

$\mathcal{T}_F := \partial F$ — a maximal monotone operator

**Optimal solution set** of (P): $X(0) = \{X \in \mathcal{X} \mid 0 \in \mathcal{T}_F(X)\}$
**ALM & PPA**

**ALM** for (D) $\iff$ **PPA** (proximal point algorithm) for (P)

The **essential objective function** of (P)

$$F(X) := \inf \{ L(w, y, S, Z, X) \mid (w, y, S, Z) \in \mathbb{R}^d \times \mathbb{R}^e \times \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \}$$

$\mathcal{T}_F := \partial F$ — a maximal monotone operator

**Optimal solution set** of (P): $X(0) = \{ X \in \mathcal{X} \mid 0 \in \mathcal{T}_F(X) \}$

Given $\sigma_k > 0$, the **proximal mapping** associated with $\sigma_k \mathcal{T}_F$:

$$\mathcal{P} := (I + \sigma_k \mathcal{T}_F)^{-1}$$

The **proximal point algorithm** (PPA):

$$X^{k+1} \approx \mathcal{P}_k(X^k), \quad \mathcal{P}_k := (I + \sigma_k \mathcal{T}_F)^{-1}$$
ALM & PPA

**ALM for (D) ⇐⇒ PPA (proximal point algorithm) for (P)**

The **essential objective function** of (P)

\[ F(X) := \inf \{ L(w, y, S, Z, X) \mid (w, y, S, Z) \in \mathbb{R}^d \times \mathbb{R}^e \times \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \} \]

\[ \mathcal{T}_F := \partial F \] — a maximal monotone operator

**Optimal solution set** of (P): \( X(0) = \{ X \in \mathcal{X} \mid 0 \in \mathcal{T}_F(X) \} \)

Given \( \sigma_k > 0 \), the **proximal mapping** associated with \( \sigma_k \mathcal{T}_F \):

\[ \mathcal{P} := (I + \sigma_k \mathcal{T}_F)^{-1} \]

The **proximal point algorithm** (PPA):

\[ X^{k+1} \approx \mathcal{P}_k(X^k), \quad \mathcal{P}_k := (I + \sigma_k \mathcal{T}_F)^{-1} \]

Criteria for approximate calculation of \( \mathcal{P}_k(X^k) \):

\[(A) \quad \|X^{k+1} - \mathcal{P}_k(X^k)\| \leq \varepsilon_k^2 / 2\sigma_k \]

\[(B) \quad \|X^{k+1} - \mathcal{P}_k(X^k)\| \leq (\eta_k^2 / 2\sigma_k)\|X^{k+1} - X^k\|^2 \]
Convergent rate

Rockafellar (1976)\textsuperscript{2} established the convergence rate under the \textbf{robust isolated calmness}

---

Rockafellar (1976)\textsuperscript{2} established the convergence rate under the robust isolated calmness.

**Robust isolated calmness:** requires the solution of problem (P) to be unique — **Restrictive**


Rockafellar (1976)\textsuperscript{2} established the convergence rate under the robust isolated calmness

**Robust isolated calmness**: requires the solution of problem (P) to be unique — Restrictive

Luque\textsuperscript{3} relaxed this with an error bound type condition


Rockafellar (1976)\textsuperscript{2} established the convergence rate under the \textit{robust isolated calmness}.

\textbf{Robust isolated calmness}: requires the solution of problem (P) to be unique — Restrictive.

Luque\textsuperscript{3} relaxed this with an \textit{error bound} type condition

- allow the multi-solution case and is satisfied when \( \mathcal{T}_F \) is \textit{polyhedral} (e.g., LASSO)

Convergent rate

Rockafellar (1976)\textsuperscript{2} established the convergence rate under the \textbf{robust isolated calmness}

\textbf{Robust isolated calmness}: requires the solution of problem (P) to be \textit{unique} — \textbf{Restrictive}

Luque\textsuperscript{3} relaxed this with an \textit{error bound} type condition

- allow the multi-solution case and is satisfied when $\mathcal{T}_F$ is \textbf{polyhedral} (e.g., LASSO)
- if $\mathcal{T}_F$ is \textbf{non-polyhedral}, it is difficult to check \textit{Luque’s condition} (counterexamples)


Convergent rate

Rockafellar (1976)\(^2\) established the convergence rate under the robust isolated calmness

**Robust isolated calmness**: requires the solution of problem (P) to be unique — Restrictive

Luque\(^3\) relaxed this with an error bound type condition

- allow the multi-solution case and is satisfied when \( \mathcal{T}_F \) is polyhedral (e.g., LASSO)
- if \( \mathcal{T}_F \) is non-polyhedral, it is difficult to check Luque’s condition (counterexamples)

**Question**: For MOPs, how to ensure the convergence of ALM and how fast it can be?

---


Theorem (asymptotic superlinear for ALM)

Assume that $T_F^{-1}(0) \neq \emptyset$. (i) Let $\{X^k\}$ be a sequence generated by the PPA with stopping criterion (A). Then $\{X^k\}$ converges to some $X^\infty \in X(0)$. (ii) If the criterion (B) is also employed and optimal solution mapping $X = T_F^{-1}$ is calm at the origin for $X^\infty$ with modulus $\kappa_p \geq 0$, then there exists $\bar{k} \geq 0$ such that for all $k \geq \bar{k}$, $\eta_k < 1$ and

$$
\text{dist}(X^{k+1}, T_F^{-1}(0)) \leq \theta_k \text{dist}(X^k, T_F^{-1}(0)),
$$

where

$$
\theta_k = (\mu_k + 2\eta_k)(1 - \eta_k)^{-1} \quad \text{with} \quad \mu_k = \kappa_p / \sqrt{\kappa_p^2 + \sigma_k^2},
$$

$$
\theta_k \to \theta_\infty = \kappa_p / \sqrt{\kappa_p^2 + \sigma_\infty^2} \quad (\theta_\infty = 0 \text{ if } \sigma_\infty = \infty).
$$

$$
\|F^*w^{k+1} + A^*y^{k+1} + S^{k+1} + Z^{k+1} - C\| \leq \tau_k^{1} \text{dist}(X^k, T_F^{-1}(0)) \quad \text{and}
$$

$$
\psi(z^{k+1}) - \psi^* \leq \tau_k^{2} \text{dist}(X^k, T_F^{-1}(0)). \quad \text{where } \tau_k^{1} \to \tau_\infty^{1} = 1/\sigma_\infty, \tau_k^{2} \to \tau_\infty^{2} = \|X^\infty\|/\sigma_\infty \quad (\tau_\infty^{1} = \tau_\infty^{2} = 0 \text{ if } \sigma_\infty = \infty).
$$
Convergence rate of ALM (cont’d)

Roughly speaking, $X^k \rightarrow X^\infty$ \textit{linearly} with a rate bounded from above by

$$\frac{\kappa_p}{\sqrt{\kappa_p^2 + \sigma_\infty^2}} < 1 \quad \text{(fast linear)}$$
Convergence rate of ALM (cont’d)

Roughly speaking, $X^k \rightarrow X^\infty$ \textbf{linearly} with a rate bounded from above by

$$\frac{\kappa_p}{\sqrt{\kappa^2_p + \sigma^2_\infty}} < 1 \quad \text{(fast linear)} \quad \text{(CG : } \frac{\sqrt{\kappa_p} - 1}{\sqrt{\kappa_p} + 1} \approx 1)$$
Convergence rate of ALM (cont’d)

Roughly speaking, \( X^k \to X^\infty \) linearly with a rate bounded from above by

\[
\frac{\kappa_p}{\sqrt{\kappa_p^2 + \sigma_\infty^2}} < 1 \quad \text{(fast linear)}
\]

\[
\text{CG : } \frac{\sqrt{\kappa_p} - 1}{\sqrt{\kappa_p} + 1} \approx 1
\]

Powell (1969): ALM ⇔ Approximate Newton method

---

Convergence rate of ALM (cont’d)

Roughly speaking, $X^k \to X^\infty$ linearly with a rate bounded from above by

$$\frac{\kappa p}{\sqrt{\kappa^2 p + \sigma_\infty^2}} < 1 \quad \text{(fast linear)}$$

(CG : $\frac{\sqrt{\kappa p} - 1}{\sqrt{\kappa p} + 1} \approx 1$)

Powell (1969)⁴: ALM $\iff$ Approximate Newton method

$$\min_{x} f(x) \quad \text{s.t. } \phi(x) = 0$$

ALM : \begin{align*}
    x^{k+1} &= \text{argmin}_x L_\sigma(x; y) \\
    y^{k+1} &= y^k + \sigma \phi(x^{k+1})
\end{align*}

Convergence rate of ALM (cont’d)

Roughly speaking, \( X^k \to X^\infty \) linearly with a rate bounded from above by

\[
\frac{\kappa_p}{\sqrt{\kappa_p^2 + \sigma_\infty^2}} < 1 \quad \text{(fast linear)}
\]

(Powell (1969): \( \kappa_p \approx 1 \))

Powell (1969): \( \text{ALM} \iff \text{Approximate Newton method} \)

\[
\begin{align*}
\min & \quad f(x) \\
\text{s.t.} & \quad \phi(x) = 0
\end{align*}
\]

ALM:

\[
\begin{align*}
x^{k+1} &= \arg\min_x L_\sigma(x; y) \\
y^{k+1} &= y^k + \sigma \phi(x^{k+1})
\end{align*}
\]

\[
y^{k+1} = y^k + \sigma \phi(x^{k+1}) \iff \text{gradient method for } \Psi_\sigma(y) := \phi(x(y, \sigma)) = 0
\]

Convergence rate of ALM (cont’d)

Roughly speaking, \( X^k \to X^\infty \) linearly with a rate bounded from above by

\[
\frac{\kappa_p}{\sqrt{\kappa_p^2 + \sigma_\infty^2}} < 1 \quad \text{(fast linear)} \quad (\text{CG} : \frac{\sqrt{\kappa_p} - 1}{\sqrt{\kappa_p} + 1} \approx 1)
\]

Powell (1969): ALM \iff \text{Approximate Newton method}

\[
\begin{align*}
\min & \quad f(x) \\
\text{s.t.} & \quad \phi(x) = 0
\end{align*}
\]

ALM:

\[
\begin{cases}
  x^{k+1} = \arg\min_x L_\sigma(x; y) \\
  y^{k+1} = y^k + \sigma \phi(x^{k+1})
\end{cases}
\]

\[y^{k+1} = y^k + \sigma \phi(x^{k+1}) \iff \text{gradient method for } \Psi_\sigma(y) := \phi(x(y, \sigma)) = 0\]

Powell (1969) shows that the Jacobian \( -J^*_\sigma \) of \( \Psi^*_\sigma \) satisfies

\[
\| -J^*_\sigma - I \| = O\left(\frac{1}{\sigma - c}\right)
\]

Convergence rate of ALM (cont’d)

Source: Cui, Sun and Toh (2018)
Augmented Lagrangian method

The convex case

The non-convex case
The non-convex case

- For the nonlinear SDP problem, Sun, Sun and Zhang (2007): ALM converges linearly under the strong regularity, i.e., $S_{\text{KKT}}$ is locally Lipschitz continuous.

---

The non-convex case

- For the nonlinear SDP problem, **Sun, Sun and Zhang (2007)**\(^5\): ALM converges linearly under the **strong regularity**, i.e., \(S_{\text{KKT}}\) is locally Lipschitz continuous.
- For the general conic optimization, **Kanzow and Steck (2018)**\(^6\): a modified ALM converges linearly under the **robust isolated calmness** of \(S_{\text{KKT}}\).

---


The non-convex case

- For the nonlinear SDP problem, Sun, Sun and Zhang (2007)\textsuperscript{5}: ALM converges linearly under the \textit{strong regularity}, i.e., $S_{\text{KKT}}$ is locally Lipschitz continuous.
- For the general conic optimization, Kanzow and Steck (2018)\textsuperscript{6}: a modified ALM converges linearly under the \textit{robust isolated calmness} of $S_{\text{KKT}}$
- A characterization for $S_{\text{KKT}}$ in general \textit{non-polyhedral} and \textit{non-convex} cases is provided by D., Sun and Zhang (2017)\textsuperscript{7}

\textsuperscript{6}C. Kanzow and D. Steck. \textit{Mathematical Programming}, 2018.
Two perturbation results
The **calmness** of the optimal solution $X$ for the convex MOPs, i.e., there exist neighborhoods $U$ of $\bar{u}$, $V$ of $\overline{X} \in X(\bar{u})$ and constant $\kappa > 0$ such that

$$X(u) \cap V \subseteq X(\bar{u}) + \kappa \|u - \bar{u}\|_B \quad \forall u \in U.$$
Calmness and strong regularity

- The **calmness** of the optimal solution $X$ for the convex MOPs, i.e., there exist neighborhoods $U$ of $\bar{u}$, $V$ of $\bar{X} \in X(\bar{u})$ and constant $\kappa > 0$ such that

  $$X(u) \cap V \subseteq X(\bar{u}) + \kappa \| u - \bar{u} \|_B \quad \forall u \in U.$$ 

- The **strong regularity** of the KKT solution $S_{KKT}$ for the general non-convex MOPs, i.e.,

  $S_{KKT}$ is a local Lipschitz continuous function.
Two perturbation results

The calmness of $X$

The strong regularity of the general MOP
Recall the following quadratic MOP and its dual:

\[(P) \quad \min \quad \frac{1}{2} \| FX - d \|^2 + \langle C, X \rangle + \theta(X)\]
\[\text{s.t.} \quad AX = b, \ X \in Q\]

\[(D) \quad \min \quad \frac{1}{2} \| w \|^2 - \langle w, d \rangle - \langle b, y \rangle + \theta^*(-S) + \delta_Q^*(-Z)\]
\[\text{s.t.} \quad F^*w + A^*y + S + Z = C\]
Recall the following quadratic MOP and its dual:

\[(P) \quad \min \quad \frac{1}{2} \|F X - d\|^2 + \langle C, X \rangle + \theta(X)\]
\[\text{s.t.} \quad AX = b, \; X \in Q\]

\[(D) \quad \min \quad \frac{1}{2} \|w\|^2 - \langle w, d \rangle - \langle b, y \rangle + \theta^*(-S) + \delta^*_Q(-Z)\]
\[\text{s.t.} \quad F^* w + A^* y + S + Z = C\]

The **Lagrange function** of (D):

\[L(w, y, S, Z, X) := \frac{1}{2} \|w\|^2 - \langle b, y \rangle - \langle d, w \rangle + \theta^*(-S) + \delta^*_Q(-Z)\]
\[+ \langle X, F^* w + A^* y + S + Z - C \rangle,\]
The convex MOPs

Recall the following quadratic MOP and its dual:

\[(P) \quad \min \quad \frac{1}{2} \|FX - d\|^2 + \langle C, X \rangle + \theta(X)\]
\[\text{s.t.} \quad AX = b, \; X \in Q\]

\[(D) \quad \min \quad \frac{1}{2} \|w\|^2 - \langle w, d \rangle - \langle b, y \rangle + \theta^*(-S) + \delta_Q^*(-Z)\]
\[\text{s.t.} \quad F^*w + A^*y + S + Z = C\]

The **Lagrange function** of \((D)\):

\[L(w, y, S, Z, X) := \frac{1}{2} \|w\|^2 - \langle b, y \rangle - \langle d, w \rangle + \theta^*(-S) + \delta_Q^*(-Z)\]
\[+ \langle X, F^*w + A^*y + S + Z - C \rangle,\]

The **essential objective function** of \((P)\)

\[F(X) := \inf \{ L(w, y, S, Z, X) \mid (w, y, S, Z) \in \mathbb{R}^d \times \mathbb{R}^e \times \mathbb{R}^{m \times n} \times \mathbb{R}^{m \times n} \}\]

\[\mathcal{T}_F := \partial F \quad \text{— a maximal monotone operator}\]

**Optimal solution set** of \((P)\): \(X(0) = \{ X \in \mathcal{X} \mid 0 \in \mathcal{T}_F(X) \}\)
The calmness of $X$

**Theorem (sufficient conditions for calmness)**

$X = T_F^{-1}$ is **calm** at $0$ for $X$ if one of the following two conditions holds:

(i) the function $\theta$ is a $C^2$-cone reducible function and the second order sufficient condition of (P) holds at $X$.

(ii) for any $v \in \partial \theta(X)$, $(\partial \theta)^{-1}$ is calm at $v$ for $X$ and there exists an optimal solution $\hat{X}$ and Lagrange multipliers $(\hat{w}, \hat{y}, \hat{S}, \hat{Z})$ such that the strict complementarity holds at $(\hat{X}, \hat{w}, \hat{y}, \hat{S}, \hat{Z})$.
The calmness of $X$

**Theorem (sufficient conditions for calmness)**

$X = T_F^{-1}$ is **calm** at 0 for $\bar{X}$ if one of the following two conditions holds:

- (i) the function $\theta$ is a $C^2$-cone reducible function and the **second order sufficient condition** of (P) holds at $\bar{X}$.

(i) the function $\theta$ is a $C^2$-cone reducible function and the **second order sufficient condition** of (P) holds at $\bar{X}$.
The calmness of $X$

**Theorem (sufficient conditions for calmness)**

$X = T_F^{-1}$ is calm at 0 for $\overline{X}$ if one of the following two conditions holds:

(i) the function $\theta$ is a $C^2$-cone reducible function and the second order sufficient condition of $(P)$ holds at $\overline{X}$.

(ii) for any $v \in \partial \theta(X)$, $(\partial \theta)^{-1}$ is calm at $v$ for $X$ and there exists an optimal solution $\hat{X}$ and Lagrange multipliers $(\hat{w}, \hat{y}, \hat{S}, \hat{Z})$ such that the strict complementarity holds at $(\hat{X}, \hat{w}, \hat{y}, \hat{S}, \hat{Z})$. 
Recall $\theta(X) = \phi(\sigma(X)) \; \forall \; X \in \mathbb{R}^{m \times n}$.

**Theorem**

Let $\phi : \mathbb{R}^m \rightarrow (-\infty, +\infty]$ be an **absolutely symmetric function**. For any $(\overline{X}, \overline{W}) \in \text{gph} \; \partial \theta$, if the inverse of the subdifferential mapping $(\partial \phi)^{-1}$ is calm at $\sigma(\overline{W})$ for $\sigma(\overline{X})$, then $(\partial \theta)^{-1}$ is calm at $\overline{W}$ for $\overline{X}$. 
Recall \( \theta(X) = \phi(\sigma(X)) \ \forall \ X \in \mathbb{R}^{m \times n} \).

**Theorem**

Let \( \phi : \mathbb{R}^m \to (-\infty, +\infty] \) be an **absolutely symmetric function**. For any \((\bar{X}, \bar{W}) \in \text{gph} \ \partial \theta\), if the inverse of the subdifferential mapping \((\partial \phi)^{-1}\) is calm at \(\sigma(\bar{W})\) for \(\sigma(\bar{X})\), then \((\partial \theta)^{-1}\) is calm at \(\bar{W}\) for \(\bar{X}\).

- \(\partial \phi\) is a **polyhedral mapping**, S.M. Robinson (1981)
Recall $\theta(X) = \phi(\sigma(X))$ $\forall X \in \mathbb{R}^{m \times n}$.

**Theorem**

Let $\phi : \mathbb{R}^m \to (-\infty, +\infty]$ be an absolutely symmetric function. For any $(\bar{X}, \bar{W}) \in \text{gph} \, \partial \theta$, if the inverse of the subdifferential mapping $(\partial \phi)^{-1}$ is calm at $\sigma(\bar{W})$ for $\sigma(\bar{X})$, then $(\partial \theta)^{-1}$ is calm at $\bar{W}$ for $\bar{X}$.

- $\partial \phi$ is a polyhedral mapping, S.M. Robinson (1981)
- $\phi$ is a piecewise linear-quadratic function, J. Sun’s Ph.D thesis (1986)
  
  e.g., $\delta_{S^n_+}$; $\delta_{\text{EDM}}$; $\| \cdot \|_2$; $\| \cdot \|_*$; $\| \cdot \|_{(k)}$; ...
Two perturbation results

The calmness of $X$

The strong regularity of the general MOP
The general MOPs

**MOP:**

\[
\min_{x \in X} \ f(x) + \theta(g(x)) \\
\text{s.t.} \quad h(x) \in Q
\]
The general MOPs

**MOP:**

\[
\min_{x \in X} f(x) + \theta(g(x)) \\
\text{s.t. } h(x) \in Q
\]

- \( \theta = \phi \circ \sigma \): a function defined on the finite dimensional matrix space
The general MOPs

MOP:

$$\min_{x \in X} f(x) + \theta(g(x))$$

s.t. $h(x) \in Q$

- $\theta = \phi \circ \sigma$: a function defined on the finite dimensional matrix space
- $\phi : \mathbb{R}^m \rightarrow (-\infty, \infty]$ is a convex piecewise linear function:

$$\phi(x) = \phi_1(x) + \phi_2(x), \quad x \in \mathbb{R}^m$$

with

$$\phi_1(x) := \max\{\langle \eta_1, x \rangle - \xi_1, \ldots, \langle \eta_l, x \rangle - \xi_l\}$$

and $\phi_2(x) := \delta_{\text{dom } \phi}(x)$ where

$$\text{dom } \phi := \{x \mid \langle \gamma_i, x \rangle - \nu_i \leq 0, \ i = 1, \ldots, r\}$$
Let \( \bar{x} \) be a feasible solution to the MOP with \( \mathcal{M}(\bar{x}) \neq \emptyset \).
Let $\bar{x}$ be a feasible solution to the MOP with $\mathcal{M}(\bar{x}) \neq \emptyset$.

The “no-gap” (strong) second order sufficient condition holds at $\bar{X}$, if for any $d \in \mathcal{C}(\bar{x}) \setminus \{0\}$ ($d \in \hat{\mathcal{C}}(\bar{x}) \setminus \{0\}$),

$$\sup_{(\bar{y}, \bar{S}) \in \mathcal{M}(\bar{x})} \left\{ \langle d, L''_{xx}(\bar{x}, \bar{y}, \bar{S})d \rangle - \Upsilon_{g(\bar{x})}(\bar{S}, g'(\bar{x})d) \right\} > 0$$
Let $\bar{x}$ be a feasible solution to the MOP with $\mathcal{M}(\bar{x}) \neq \emptyset$.

The “no-gap” \textbf{(strong) second order sufficient condition} holds at $\bar{X}$, if for any $d \in \mathcal{C}(\bar{x}) \setminus \{0\}$ ($d \in \hat{\mathcal{C}}(\bar{x}) \setminus \{0\}$),

$$\sup_{(\bar{y}, \bar{S}) \in \mathcal{M}(\bar{x})} \left\{ \langle d, L''_{xx}(\bar{x}, \bar{y}, \bar{S})d \rangle - \gamma g(\bar{x}) (\bar{S}, g'(\bar{x})d) \right\} > 0$$

- $\mathcal{C}(\bar{x})$: the \textbf{critical cone}
Let $\bar{x}$ be a feasible solution to the MOP with $\mathcal{M}(\bar{x}) \neq \emptyset$.

The "no-gap" (strong) second order sufficient condition holds at $\bar{X}$, if for any $d \in C(\bar{x}) \setminus \{0\}$ ($d \in \hat{C}(\bar{x}) \setminus \{0\}$),

$$\sup_{(\bar{y}, \bar{S}) \in \mathcal{M}(\bar{x})} \left\{ \langle d, L''_{xx}(\bar{x}, \bar{y}, \bar{S})d \rangle - \Upsilon_{g(\bar{x})}(\bar{S}, g'(\bar{x})d) \right\} > 0$$

- $C(\bar{x})$: the critical cone
- $\hat{C}(\bar{x}) := \bigcap_{(\bar{y}, \bar{S}) \in \mathcal{M}(\bar{x})} \text{app}(\bar{y}, \bar{S})$ with $\text{app}(\bar{y}, \bar{S})$ is an approximation set of $\text{aff}(C(\bar{x}))$
Let $\bar{x}$ be a feasible solution to the MOP with $\mathcal{M}(\bar{x}) \neq \emptyset$.

The "no-gap" (strong) second order sufficient condition holds at $\bar{X}$, if for any $d \in \mathcal{C}(\bar{x}) \setminus \{0\}$ ($d \in \hat{\mathcal{C}}(\bar{x}) \setminus \{0\}$),

\[
\sup_{(\bar{y}, \bar{S}) \in \mathcal{M}(\bar{x})} \left\{ \langle d, L''_{xx}(\bar{x}, \bar{y}, \bar{S})d \rangle - \gamma_{g(\bar{x})}(\bar{S}, g'(\bar{x})d) \right\} > 0
\]

- $\mathcal{C}(\bar{x})$: the critical cone
- $\hat{\mathcal{C}}(\bar{x}) := \bigcap_{(y, S) \in \mathcal{M}(\bar{x})} \text{app}(y, S)$ with $\text{app}(y, S)$ is an approximation set of $\text{aff}(\mathcal{C}(\bar{x}))$
- $\gamma_{g(\bar{x})}(\bar{S}, g'(\bar{x})d)$: the so-called "$\sigma$-term"
The constraint nondegeneracy for the MOP:

\[
\begin{bmatrix}
    h'(\bar{x}) \\
g'(\bar{x})
\end{bmatrix} \mathbb{X} + \begin{bmatrix}
    \text{lin} \left( T_Q(h(\bar{x})) \right) \\
    T^{\text{lin}}(g(\bar{x}))
\end{bmatrix} = \begin{bmatrix}
    \mathbb{R}^d \\
    \mathbb{V}
\end{bmatrix},
\]

where the linear space $T^{\text{lin}}$ is called the linearity space of the $\theta$ defined by

\[
T^{\text{lin}}(Z) := \{ H \mid \theta'(Z; H) = -\theta'(Z; -H) \}.
\]
The **constraint nondegeneracy** for the MOP:

\[
\begin{bmatrix}
  h'(\bar{x}) \\
g'(\bar{x})
\end{bmatrix} \mathbb{X} + \begin{bmatrix}
  \text{lin} (T_Q(h(\bar{x}))) \\
  T^\text{lin}(g(\bar{x}))
\end{bmatrix} = \begin{bmatrix}
  \mathbb{R}^d \\
  \mathbb{V}
\end{bmatrix},
\]

where the linear space $T^\text{lin}$ is called the linearity space of the $\theta$ defined by

\[
T^\text{lin}(Z) := \{H \mid \theta'(Z; H) = -\theta'(Z; -H)\}
\]

Most importantly, the **explicit formulas** for above are now available for the case $\phi$ is a convex piecewise linear function *Cui and Ding (2018)*.
Theorem

Let $\bar{x} \in X$ be a feasible solution to the MOP with $\mathcal{M}(\bar{x}) \neq \emptyset$. Suppose that the Robinson constraint qualification (RCQ) holds at $\bar{x}$ and $(\bar{y}, \bar{S}) \in \mathcal{M}(\bar{x})$. Then, the following statements are equivalent:

(i) The strong second order sufficient condition and constraint nondegeneracy for the MOP both hold at $\bar{x}$;

(ii) $(\bar{x}, \bar{y}, \bar{S})$ is a strongly regular solution of the KKT equations.
MOPs in Data Science
Conclusions

**MOPs** in Data Science

**ALM** for MOPs
Conclusions

**MOPs** in Data Science

**ALM** for MOPs

Sufficient conditions for the *calmness* of $X$
Conclusions

**MOPs** in Data Science

**ALM** for MOPs

Sufficient conditions for the *calmness* of $X$

Characterization of the *strong regularity* of the general MOP
Thank you!