Linear Convergence of Alternating Direction Methods

Wotao Yin

(joint work with Wei Deng)

Department of Computational and Applied Mathematics
Rice University

MIIS - July 9, 2012
Problem

Model:

\[
\min_{x,y} \quad f(x) + g(y)
\]

s.t. \quad Ax + By = b.

- \(f(x)\) and \(g(y)\) are convex functions, maybe nonsmooth
Yin Zhang: “the idea of ADM goes back to Sun-Tze (400 BC) and Caesar (100 BC):

“Divide and Conquer.” — Julius Caesar (100-44 BC) 
“远交近攻”, “各个击破”. — Sun-Tzu (400 BC) 

Back to 1950s–70s, it appears as operator splitting for PDEs and studied by Douglas, Peaceman, and Rachford, and then Glowinsky et al.’81-89, Gabay’83.

Subsequent studies are in the context of variational inequality (Eckstein and Bertsekes’92, He et al.’02)

Reappeared as Split Bregman (Goldstein and Osher)

Extensions to multiple blocks (He, Yuan, and collaborators)
Alternating Direction Method (ADM)

Augmented Lagrangian:

\[
\mathcal{L}(x, y, \lambda) = f(x) + g(y) - \lambda^T(Ax + By - b) + \frac{\beta}{2} \|Ax + By - b\|^2.
\]

**Algorithm 1:** the classic Alternating Direction Method (ADM)

1. Initialize \( x^0, \lambda^0, \beta > 0; \)
2. for \( k = 0, 1, \ldots \) do
   3. \( y^{k+1} = \text{arg min}_y \mathcal{L}_A(x^k, y, \lambda^k); \)
   4. \( x^{k+1} = \text{arg min}_x \mathcal{L}_A(x, y^{k+1}, \lambda^k); \)
   5. \( \lambda^{k+1} = \lambda^k - \beta(Ax^{k+1} + By^{k+1} - b). \)
Algorithm 2:

1. Choose $Q \succeq 0$ and a symmetric matrix $P$. Initialize $x^0, \lambda^0, \beta > 0, \gamma > 0$;
2. for $k = 0, 1, \ldots$ do
3. $y^{k+1} = \text{arg min}_y \mathcal{L}_A(x^k, y, \lambda^k) + \frac{1}{2}(y - y^k)^T Q (y - y^k)$;
4. $x^{k+1} = \text{arg min}_x \mathcal{L}_A(x, y^{k+1}, \lambda^k) + \frac{1}{2}(x - x^k)^T P (x - x^k)$;
5. $\lambda^{k+1} = \lambda^k - \gamma \beta (Ax^{k+1} + By^{k+1} - b)$.

Goals of choosing $P$ and $Q$:
- To make subproblems much easier to solve
- To reduce total solution time

Step size $\gamma$ depends on choices $P$ and $Q$. Less exact subproblems requires smaller $\gamma$.

Allows indefinite $P$, requiring $\gamma < 1$.

Related to
- [He, Liao, Han, and Yang, 2002] uses $\|Ax + By - b\|^2_{H_k}$, proves convergence for $\gamma = 1$ and differentiable $f$ and $g$
- [Zhang, Burger, and Osher, 2011] replaces $\gamma$ by $C \succ 0$, proves convergence for $\|C\| \leq 1$ and $A = I$
Convergence results

▶ [Goldfarb and Ma, 2009] Assume: a Jacobi-type ADM, \( f, g \) smooth and \( \nabla f, \nabla g \) Lipschitz
Result: objective \( \sim O(1/k) \); accelerated alg, objective \( \sim O(1/k^2) \)

▶ [Goldfarb, Ma, and Scheinberg, 2009] Assume: (Gauss-Seidel) ADM, \( f \) smooth and \( \nabla f \) Lipschitz
Result: same as above

▶ [He and Yuan, 2012] Assume: at least one subproblem is exactly solved
Result: let \( u^k := (x^k, y^k, \lambda^k) \),
  - ergodic: objective error+\((\tilde{u}^k - u^*)^T F(u^*) \sim O(1/k)\)
  - non-ergodic \( \|u^k - u^{k+1}\| \sim O(1/k) \)

▶ [Goldstein et al., May 2012] Assume: \( f, g \) strongly convex and \( \nabla f, \nabla g \) Lipschitz, \( g \) quadratic
Result: dual objective \( \sim O(1/k^2) \)

▶ We assume: \( f \) strongly convex, \( \nabla f \) Lipschitz, rank conds on \( A \) and \( B \)
Result: points and objective \( \sim O(1/c^k) \), where \( c > 1 \)
Application: consensus optimization

Consider a network of $N$ nodes and model

$$\min_x \sum_{i=1}^{N} f_i(x).$$

To reach consensus on solution $x$, introduce *local copies* $x_i$, $i = 1, \ldots, N$.

Global consensus model

$$\min_{\{x_i\}, y} \sum_{i=1}^{N} f_i(x_i), \quad \text{s.t. } x_i - y = 0, \ i = 1, \ldots, N.$$  

Decentralized consensus model

$$\min_{\{x_i\}, y} \sum_{i=1}^{N} f_i(x_i), \quad \text{s.t. } x_i - y_{ij} = 0, \ \forall \ \text{neighbor} \ (i, j).$$
Application: regularization

Regularization model:

$$\min_y f(By - b) + g(y)$$

Many examples in statistics, machine learning, inverse problems, imaging, compressive sensing, ...

- $By = b$ is often *ill-conditioned*: noise in $b$, non-unique solutions
- $f(\cdot)$ smooth / $\min f(By - b)$ gives $By = b$ or $By \approx b$
- $\min g(y)$ looks for $y$ with *wanted properties*
  - Examples: harmonic, low energy, sparse, low rank, sharp edges, ...
- $g(\cdot)$: may or may not be smooth
- *Difficult to minimize both $f$ and $g$ the same time*
Application: regularization

Regularization model:

$$\min_y f(By - b) + g(y)$$

Many examples in statistics, machine learning, inverse problems, imaging, compressive sensing, ...

- $By = b$ is often *ill-conditioned*: noise in $b$, non-unique solutions
- $f(\cdot)$ smooth / $\min f(By - b)$ gives $By = b$ or $By \approx b$
- $\min g(y)$ looks for $y$ with *wanted properties*
  
  Examples: harmonic, low energy, sparse, low rank, sharp edges, ...
- $g(\cdot)$: may or may not be smooth
- *Difficult to minimize both $f$ and $g$ the same time*

ADM:

$$\min f(x) + g(y), \text{ s.t. } x + By = b.$$
Inexact ADM: I. Prox-Linear

- When a subproblem, say, the $x$-subproblem
  \[
  \min_{x \in \mathcal{X}} f(x) + \frac{\beta}{2} \|Ax + By^{k+1} - b - \lambda^k / \beta\|^2
  \]
  is not easy to solve because of $A$:
- Apply
  \[
  P = \frac{\beta}{\tau} I - \beta A^T A
  \]
- New: prox-linearize $x$-subproblem
  \[
  x^{k+1} \leftarrow \min_{x \in \mathcal{X}} f(x) + \beta \left( (g^k)^T (x - x^k) + \frac{1}{2\tau} \|x - x^k\|^2 \right),
  \]
  where $\tau > 0$ is a proximal parameter, and
  \[
  g^k := A^T(Ax^k + By^{k+1} - b - \lambda^k / \beta).
  \]

Existing work: [Chen and Teboulle, 1993], [Tao and Yuan, 2010], [Yang and Zhang, 2011]
Sparse optimization

Model
\[
\min_y \frac{1}{2} \|By - b\|_2^2 + \mu \|y\|_1.
\]

ADM
\[
\min_{x,y} \frac{1}{2} \|x\|_2^2 + \mu \|y\|_1 \text{ s.t. } x + By = b.
\]

- x-subproblem: closed form.
- y-subproblem: use \( Q = \eta I - \beta B^T B \succeq 0 \)
\[
y^{k+1} \leftarrow \arg \min_y \mu \|y\|_1 - \langle B\lambda^k + \beta B^T B y^k, y \rangle + \frac{\eta}{2} \|y - y^k\|_2^2
\]
reduces to \( \min_{y_i} \) for each \( i \). also closed form using soft-thresholding.
Low-rank matrix recovery

Model [Fazel, 2002, Candes and Recht, 2009, Recht et al., 2010]

\[
\min_Y \frac{1}{2} \| B(Y) - b \|_2^2 + \mu \| Y \|_*
\]

where \( B(Y) \) has linear measurements of \( Y \) and \( \| Y \|_* \) is the sum of singular values of \( Y \).

ADM

\[
\min_{x, Y} \frac{1}{2} \| x \|_2^2 + \mu \| Y \|_* \text{ s.t. } x + B(Y) = b.
\]

\( x \)-subproblem: closed form.

\( y \)-subproblem: closed form using \( Q \) and singular value thresholding
Matrix decomposition to *low-rank + sparse*

Model [Candès et al., 2009]

\[
\min_{L,S} \frac{1}{2} \| A(L) + B(S) - M \|_2^2 + \mu_1 \|L\|_* + \mu_2 \|S\|_1
\]

*L* is expected to be low rank. *S* is expected to contain few nonzeros.

ADM

\[
\min_{x,Y} \frac{1}{2} \|x\|_2^2 + \mu_1 \|L\|_* + \mu_2 \|S\|_1, \text{ s.t. } x + A(L) + B(S) = M.
\]

*x*-subproblem: *closed form*.

*(L, S)*-subproblem: not easy without Q; but *closed form* with proper Q

▶ *singular value thresholding* on *L*

▶ *soft thresholding* on *S*

Also by [Ma et al., 2012] for inverse covariance matrix recovery
One-step projected gradient descent:

\[
x^{k+1} = \mathcal{P}_\mathcal{X} \left( x^k - \alpha^k g^k \right)
\]

where

\[
g^k = \nabla f(x^k) + \beta A^T (Ax^k + By^{k+1} - b - \lambda^k / \beta).
\]

Also see [Chen, Hager, Yashtini, and Ye, 2012] for BB/safeguard step sizes.
Inexact ADM: III. Approximating $A^T A$

- In $x$-subproblem: $\frac{\beta}{2} \|Ax + By - b\|_2^2$ contains $\frac{\beta}{2} x^T A^T Ax$
- Suppose $A^T A \approx D$ and $D$ is easier.
  Example: $A$: non-regular Fourier operation; $D$: fast Fourier transform (FFT)
- Apply:
  \[ P = \beta (D - A^T A). \]
- Then
  \[ \frac{\beta}{2} \|Ax + By - b\|_2^2 + \frac{\beta}{2} x - x^k \|_P^2 = \frac{\beta}{2} x^T Dx + \mathcal{L}(x) + C. \]
  where $\mathcal{L}(x)$ is linear in $x$. 
Other examples

- Medical imaging problems
- Regularized regression problems
- Quadratic semi-definite programming
- Optimization problems on networks
  many more......
Convergence

Extending from [He, Liao, Han, and Yang, 2002, Zhang, Burger, and Osher, 2011]:

**Theorem**

Assume \( f \) and \( g \) are convex, a KKT point exists, and \( Q = 0 \) or \( Q \succ 0 \).

If \( \gamma \) obeys

1. \((2 - \gamma) \succ (\gamma - 1) \beta A^T A, \) if \( P \neq 0 \);
2. \( 0 < \gamma < \frac{1}{2}(\sqrt{5} + 1), \) if \( P = 0 \),

then \( \exists \) a KKT point \( u^* = (x^*, y^*, \lambda^*) \) such that \( \|u^k - u^*\|_G \to 0 \) where

\[
G = \begin{bmatrix}
P + \beta A^T A \\
Q \\
\frac{1}{\beta \gamma} I
\end{bmatrix} \succ 0.
\]

**Comments:**

1. Need \( u^k = (x^k, y^k, \lambda^k) \) be bounded, but easy to have.
2. Condition 1 gives
   - \( \tau \|A\|^2 + \gamma < 2 \) for prox-linear \( P \) [Yang and Zhang, 2011];
   - \( \beta \|A\|^2/(\alpha^{-1} - \|H_f\|) + \gamma < 2 \) for gradient-descent \( P \);
   - \( \gamma < \bar{\gamma} < 1 \) for indefinite \( P \).
Scenarios for global linear convergence

<table>
<thead>
<tr>
<th>#</th>
<th>strongly convex</th>
<th>Lipschitz continuous</th>
<th>full row rank</th>
<th>remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$f$</td>
<td>$\nabla f$</td>
<td>$A$</td>
<td>if $Q \succ 0$, $B$ has full column rank</td>
</tr>
<tr>
<td>2</td>
<td>$f, g$</td>
<td>$\nabla f$</td>
<td>$A$</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>$f$</td>
<td>$\nabla f, \nabla g$</td>
<td>$[A, B]$</td>
<td>$B$ has full column rank</td>
</tr>
<tr>
<td>4</td>
<td>$f, g$</td>
<td>$\nabla f, \nabla g$</td>
<td>$[A, B]$</td>
<td>-</td>
</tr>
</tbody>
</table>

Comments:
- Scenarios 1 & 3 are more common than 2 & 4
- Without Lipschitz $\nabla g$, full row-rank of $A$ is needed
- Full row-rank of $[A, B]$ is a weak condition; almost always satisfied
Summary of Linear Convergent Quantities

**Definition:** The convergence of \( u^k \rightarrow u^* \) is

- **Q-linear**, if \( \exists \mu \in (0, 1) \) such that \( \frac{\|u^{k+1} - u^*\|}{\|u^k - u^*\|} \leq \mu \);

- **R-linear**, if \( \exists \) Q-linearly convergent \( \{\sigma^k\} \) such that \( \|u^k - u^*\| \leq \sigma^k \).
**Summary of Linear Convergent Quantities**

**Definition:** The convergence of \( u^k \to u^* \) is

- **Q-linear**, if \( \exists \mu \in (0, 1) \) such that \( \frac{\|u^{k+1} - u^*\|}{\|u^k - u^*\|} \leq \mu \);

- **R-linear**, if \( \exists \) Q-linearly convergent \( \{\sigma^k\} \) such that \( \|u^k - u^*\| \leq \sigma^k \).

<table>
<thead>
<tr>
<th>case</th>
<th>( P, \hat{P} )</th>
<th>( Q )</th>
<th>any scenario 1 – 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q-linear convergence</td>
<td>R-linear convergence*</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>( P = 0 )</td>
<td>= 0</td>
<td>( (Ax^k, \lambda^k) )</td>
</tr>
<tr>
<td>2</td>
<td>( \hat{P} \succ 0 )</td>
<td>= 0</td>
<td>( (x^k, \lambda^k) )</td>
</tr>
<tr>
<td>3</td>
<td>( P = 0 )</td>
<td>( \succ 0 )</td>
<td>( (Ax^k, y^k, \lambda^k) )</td>
</tr>
<tr>
<td>4</td>
<td>( \hat{P} \succ 0 )</td>
<td>( \succ 0 )</td>
<td>( (x^k, y^k, \lambda^k) )</td>
</tr>
</tbody>
</table>

* In cases 1 and 2, scenario 1, R-linear convergence \( y^k \) requires full column rank of \( B \); otherwise, only \( By^k \) has R-linear convergence.
Rate of Convergence

Scenario 1 and \( P = Q = 0 \):

\[
\mu = \left( 1 + 2\nu_f / \left( \beta \| A \|^2 + \frac{L_f^2}{\beta \lambda_{\min}(AA^T)} \right) \right)^{-1} \in (0, 1).
\]

To minimize rate, set \( \beta = \frac{L_f}{\| A \| \sqrt{\lambda_{\min}(AA^T)}} \) and obtain

\[
\mu_{\min} = \left( 1 + \frac{1}{\kappa_f \kappa_A} \right)^{-1}.
\]

Notation:

- \( L_f \): Lipschitz constant of \( \nabla f \);
- \( \nu_f \): strong convexity modulus of \( f \);
- \( \kappa_f = L_f / \nu_f \): condition number of \( f \);
- \( \kappa_A = \sqrt{\sigma_{\max}(AA^T) / \sigma_{\min}(AA^T)} \): condition number of \( A \).


