Differential Inclusions for Modeling Nonsmooth ADMM Variants: A Continuous Limit Theory

Huizhuo Yuan\(^1\), Yuren Zhou\(^2\), Chris Junchi Li\(^3\), Qingyun Sun\(^4\)

\(^1\)Peking University, \(^2\)Duke University, \(^3\)Tencent AI Lab, \(^4\)Stanford University

Poster: Wed Jun 12, 2019@Pacific Ballroom 210

ICML 2019, Long Beach, CA
Goal is to solve

\[ \minimize_{x \in \mathbb{R}^d} f(x) + g(Ax) \]  

\[ (1) \]

- \( f(x) \) and \( g(y) \) are defined on \( \mathbb{R}^d \) and \( \mathbb{R}^m \), separately
- Allow \( f \) and \( g \) to be nonsmooth functions in (1)

Rewrite (1) as

\[ \minimize_{x \in \mathbb{R}^d, z \in \mathbb{R}^m} f(x) + g(z) \]

\[ \text{subject to } Ax - z = 0 \]  

Scatters everywhere in statistical learning and signal processing:
Lasso, logistic regression, elastic net, and many more
Alternating Direction Method of Multipliers (ADMM)

\[
\text{minimize} \quad f(x) + g(z)
\]
\[
\text{subject to} \quad Ax - z = 0
\]

- We adopt the Generalized ADMM (G-ADMM) setting for solving (2), which introduces a new relaxation parameter \( \alpha \in (0, 2) \)

Algorithm proposed by [Eckstein-Bertsekas 1992]:

\[
x_{k+1} = \arg\min_x \left\{ f(x) + \frac{\rho}{2} \|Ax - z_k + u_k\|^2 \right\} \quad (3a)
\]

\[
z_{k+1} = \arg\min_z \left\{ g(z) + \frac{\rho}{2} \left\| \alpha Ax_{k+1} + (1 - \alpha) z_k - z + u_k \right\|^2 \right\} \quad (3b)
\]

\[
u_{k+1} = u_k + (\alpha Ax_{k+1} + (1 - \alpha) z_k - z_{k+1}) \quad (3c)
\]

- When \( \alpha = 1 \), convergence rate is known
Linearized ADMM

- $f$ is nonsmooth with easy proximal mappings
- First-order Taylor approximation to the second term of (3a):

$$x_{k+1} = \arg\min_x \left\{ f(x) + \frac{\tau L}{2} \left\| x - \left( x_k - \rho \frac{\tau L}{T} A^\top (Ax_k - z_k + u_k) \right) \right\|_2^2 \right\}$$  (4a)

- Total variation minimization problem [Ruding-Osher-Fatemi 1992]

$$\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| x - b \|_2^2 + \lambda \| z \|_1 \\
\text{subject to} & \quad z = Dx
\end{align*}$$  (4a) and (3b) respectively correspond to the proximal mappings of $\| \cdot \|_1$ and $\frac{1}{2} \| \cdot - b \|_2^2$
Gradient Based ADMM

- $f$ is differentiable but does not have an easy proximal mapping
- $g$ is nonsmooth with easy proximal mappings
- A gradient step is taken instead of minimizing the augmented Lagrangian function directly

$$ x_{k+1} = x_k - \frac{1}{\tau_G} \left( \nabla f(x_k) + \rho A^\top (Ax_k - z_k + u_k) \right) $$ (5a)

- Sparse logistic regression problem as an example

$$ \min_{x} \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-b_i(a_i^\top x + v))) + \lambda \|x\|_1 $$ (6)
We study the continuous limit of the generalized ADMM (G-ADMM):

- The seminal work [Su-Boyd-Candes 2014] provide new insights on understanding the convergence of (accelerated) gradient method: connecting (a second-order) ODE to the continuous limit of AGM


- Very recently, [Franca-Robinson-Vidal 2018a, b] made a significant step towards understanding G-ADMM using the tools of Differential Equation for the cases where $f$ and $g$ are both smooth

- We now extend the analysis to problems with nonsmooth $f$ and $g$, using Differential Inclusion
Continuous Limit of Linearized and Gradient Based ADMM

The continuous-time limit of the iterates \( \{x_k\} \) of linearized ADMM (4a) and gradient-based ADMM (5) is given by the differential inclusion

\[
0 \in \partial F(X(t)) + \left( cI + \frac{1 - \alpha}{\alpha} A^\top A \right) \dot{X}(t)
\]  

(7)

Solution \( X(t) \) of differential inclusion (7) has \( O(t^{-1}) \) convergence rate:

\[
F(X(t)) - F(x^*) \leq \frac{\kappa_1^2 \|x_0 - x^*\|_2^2}{2t}
\]

- Rescale the time by setting \( t = \rho^{-1} k \)
- \( \rho \to \infty \) and \( \tau_L/\rho \to c \in (0, \infty) \) (\( \tau_G/\rho \to c \) for gradient-based)
- Initial value \( X(0) = x_0 \)
- \( \kappa_1^2 \) is defined to be the largest eigenvalue of \( \left( cI + (1 - \alpha)/\alpha A^\top A \right) \)
Continuous Limit of G-ADMM

The continuous limit of iterates of \( \{x_k\} \) in Algorithm (3) is given by the following differential inclusion:

\[
\frac{1}{\alpha} (A^\top A) \dot{X}(t) + \partial F(X(t)) \ni 0
\]  

(8)

Let \( x^* \) be a minimizer of \( F \). Solution \( X(t) \) of differential inclusion (8) has \( O(t^{-1}) \) convergence rate:

\[
F(X(t)) - F(x^*) \leq \frac{\sigma_1^2 \|x_0 - x^*\|^2}{2\alpha t}
\]  

(9)

- Rescale the time by setting \( t = \rho^{-1} k \)
- \( \rho \rightarrow \infty \)
- Initial value \( X(0) = x_0 \)
- \( \sigma_1 \) is defined to be the largest singular value of matrix \( A \)
Continuous Limit of Accelerated G-ADMM

\[
\frac{1}{\alpha} (A^\top A) \left( \ddot{X}(t) + \frac{r}{t} \dot{X}(t) \right) + \partial F(X(t)) \ni 0
\] (10)

- Algorithm (omitted here) first proposed by [Goldstein-O’Donoghue-Setzer-Baraniuk 2014]

Theorem

- (High Friction) When \( r \geq 3 \)

\[
F(X(t)) - F(x^*) \leq \frac{C(r, \alpha, \sigma_1) \|x_0 - x^*\|^2}{t^2}
\]

- (Low Friction) When \( 0 < r < 3 \)

\[
F(X(t)) - F(x^*) \leq \frac{C(r, \alpha, \sigma_1) \|x_0 - x^*\|^2}{t^{2r/3}}
\]
A Numerical Example

Total Variation Minimization: Numerical Experiments

minimize \( x, z \in \mathbb{R}^n \)
\[
\frac{1}{2} \| x - b \|_2^2 + \lambda \| z \|_1
\]
subject to \( z = Dx \)

Fits to our problem with \( A = D \), \( f(x) = \frac{1}{2} \| x - b \|_2^2 \) and \( g(z) = \lambda \| z \|_1 \)

**Figure:** On total variation minimization problem, the plots are the trajectory of linearized ADMM with \( \rho = 10 \) and the corresponding differential inclusion, the first plot is for different \( \alpha \) from 2 to 3 when \( c = 10 \), second plot is for different \( c \) from 1 to 32 when \( \alpha = 1.6 \)
A Numerical Example

Sparse Logistic Regression: Numerical Experiments

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{i=1}^{N} \log(1 + \exp(-b_i(a_i^\top x + v))) + \lambda \|z\|_1 \\
\text{subject to} & \quad z = x
\end{align*}
\]

Fits to our problem with \( \bar{x} = (x, v) \), \( f(\bar{x}) = \log(1 + \exp(-b_i(a_i^\top x + v))) \), \( A = I \), and \( g(\bar{x}) = \lambda \|x_{1:n}\|_1 \)

**Figure:** On sparse logistic regression, the plots are gradient ADMM and the differential inclusion when \( \rho = 10 \), first plot is for different \( \alpha \) from \( 2^{-3} \) to \( 2 \) when \( c = 10 \), second plot is for different \( c \) from 1 to 32 when \( \alpha = 1.6 \).
Conclusion

- ADMM is a very popular practical algorithm for large-scale statistical learning and signal processing tasks.
- Differential inclusions associated with nonsmooth ADMM variants can provide new insights into those algorithms.
- We provide the first formulation of those differential inclusions for G-ADMM with relaxation parameters.
- Continuous-time rate in (9) matches existing discrete-time analysis [He-Yuan 2012, Eckstein-Yao 2015], but can be proved sharper than $O(t^{-1})$. 

Huizhuo Yuan, Yuren Zhou, Chris Junchi Li, Q
Thank You!

Poster: Wed Jun 12, 2019@Pacific Ballroom 210