A Three-Operator Splitting Scheme and its Applications

Wotao Yin (UCLA Math)

joint with Damek Davis

ICCM'16 — Beijing
Background
What is “splitting”? 

- Sun-Tzu: “远交近攻”, “各个击破” (400 BC)
- Caesar: “divide-n-conquer” (100–44 BC)
- Principle of computing: reduce a problem to simpler subproblems
- Example: find $x \in C_1 \cap C_2$ by alternatively projecting to $C_1$ and $C_2$
Some basic principles of splitting

split:

- $x/y$ directions
- linear from nonlinear
- smooth from nonsmooth
- spectral from spatial
- convection from diffusion
- composite operators

\[(I - \lambda(A + B))^{-1} \text{ to } (I - \lambda A)^{-1} \text{ and } (I - \lambda B)^{-1}\]

Also

- domain decomposition
- block-coordinate descent
- column generation, Bender’s decomposition
Operator splitting pipeline

1. Formulate into the inclusion problem

\[ 0 \in A(x) + B(x) \]

where \( A \) and \( B \) are operators, possibly set-valued

2. Apply \textbf{operator splitting} to obtain a fixed-point operator \( T \):

\[ z^{k+1} \leftarrow Tz^k \]

3. Efficiency requires: computing \( T \) reduces to computing \( A \) and \( B \) separately

4. Convergence requires:
   - fixed-point \( z^* = Tz^* \) recovers a solution \( x^* \)
   - \( T \) is “\textbf{contractive}” or, more weakly, “\textbf{averaged}”
Example: constrained minimization

- \( C \) is a convex set. \( f \) is a differentiable convex function.

\[
\begin{aligned}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in C
\end{aligned}
\]

- equivalent inclusion problem:

\[
0 \in N_C(x) + \nabla f(x)
\]

- projected gradient method:

\[
x^{k+1} \leftarrow \text{proj}_C \circ \left( I - \gamma \nabla f \right) x^k
\]
Example: Jacobi parallel ADMM

\[
\begin{align*}
\text{minimize} & \quad f_1(x_1) + \cdots + f_m(x_m) + g(x) \\
\text{subject to} & \quad A_1 x_1 + \cdots + A_m x_m = b.
\end{align*}
\]

- **require**: convex \( f_i \) (nonsmooth ok); convex and smooth \( g \); linear \( A_i \)

- **examples**: LP, QP, signal processing, control, exchange problems, \ldots

- **equivalent condition**: with dual variable \( z \),

\[
0 \in \begin{bmatrix}
\partial f_1 & -A_1^T \\
\vdots & \vdots \\
\partial f_m & -A_m^T \\
A_1 & \cdots & A_m & 0
\end{bmatrix}\begin{bmatrix}
x_1 \\
\vdots \\
x_m \\
z
\end{bmatrix} + \begin{bmatrix}
\nabla_1 g(x) \\
\vdots \\
\nabla_m g(x) \\
b
\end{bmatrix}
\]
(skipping about two steps) **forward-backward splitting** gives the algorithm

\[
x_{i}^{k+1} = \arg \min_{x_i} f_i(x_i) + \langle \nabla_i g(x_i^k) - A_i^T z^k + \sigma A_i^T (Ax_i^k - b), x_i \rangle + \frac{1}{2} \| x_i - x_i^k \|_2^2
\]

\forall i = 1, \ldots, m \text{ in parallel}

\[
z^{k+1} = z^k - \sigma (Ax^k - b)
\]

where \( Ax = \sum_{i=1}^{m} A_i x_i \)

**why nice?**

- solve \( x_i \)-subproblems and update \( z \) in parallel
- all \( f_i \), all \( A_i \) and \( g \) are treated separately
Convergence basics
Contractive operator

- **definition:** $T$ is contractive if, for some $L \in [0, 1)$,

\[ ||Tx - Ty|| \leq L||x - y||, \quad \forall x, y \]
Banach fixed-point theorem

- **Theorem:** If $T$ is contractive, then
  - $T$ has a unique fixed-point $x^*$ (existence, uniqueness)
  - $\|x^k - x^*\| \leq L^k \|x^0 - x^*\| \to 0$ (convergence, speed)

- Holds in a Banach space

- Also known as the Picard-Lindelöf Theorem
Between $L = 1$ and $L < 1$

- $L < 1$: geometric convergence
- $L = 1$: iterates are bounded, but may diverge
- Convergence of many algorithms **cannot** be characterized by $L$
  - Alternative projection (von Neumann)
  - Gradient descent
  - Proximal-point algorithm
  - Operator splitting algorithms
Averaged operator

- **residual operator:** \( R := I - T \). Hence, \( Rx^* = 0 \iff x^* = Tx^* \)

- **averaged operator:** from some \( \eta > 0 \),

\[
\|Tx - Ty\|^2 \leq \|x - y\|^2 - \eta \|Rx - Ry\|^2, \quad \forall x, y
\]

- **story:** set \( y \) as a fixed point, then distance to \( y \) improve by the amount of fixed-point residual

- **property\(^1\):** if \( T \) has a fixed point, then \( x^{k+1} \leftarrow Tx^k \) converges weakly to a fixed point

\(^1\) Krasnosel’skiǐ’57, Mann’56
Why called “averaged”?

Lemma

For $\alpha \in (0, 1)$, $T$ is $\alpha$-averaged if, and only if, there exists a nonexpansive (1-Lipschitz) map $T'$ so that

$$T = (1 - \alpha)I + \alpha T'.$$
Composition of averaged operators

Useful Theorem:

\[ T_1, T_2 \text{ nonexpansive } \Rightarrow \ T_1 \circ T_2 \text{ nonexpansive} \]

\[ T_1, T_2 \text{ averaged } \Rightarrow \ T_1 \circ T_2 \text{ averaged} \]

(note: the averagedness constant \( \alpha \) gets worse.)
How to get an averaged-operator composition?
Monotone operators

- **definition:** $A : \mathcal{H} \to \mathcal{H}$ (single- or set-valued) is **monotone** if

\[
\langle Ax - Ay, x - y \rangle \geq 0, \quad \forall x, y
\]

(if set-valued, inequality holds for each entry)

- **have many examples:**
  - subdifferential $\partial f$ of a convex function $f$
  - symmetric positive semidefinite linear operator
  - skew-symmetric linear operator
Forward-backward splitting

- **require:** \( A \) monotone, \( B \) monotone and single-valued

- **derive:**

\[
0 \in Ax + Bx \iff x - Bx \in x + Ax \\
\iff (I - B)x \in (I + A)x \\
\iff (I + A)^{-1}(I - B)x = x
\]

\[
\begin{array}{c}
\text{backward} \\
\text{forward} \\
\text{operator} \quad T_{\text{FBS}}
\end{array}
\]

- Although \((I + A)\) may be set-valued, \((I + A)^{-1}\) is single-valued!
- **forward-backward splitting (FBS) operator** (Mercier’79): for $\gamma > 0$

$$T_{\text{FBS}} := (I + \gamma A)^{-1} \circ (I - \gamma B)$$

- **key properties:**
  - if $B$ is $\beta$-cocoercive\(^2\) and $\gamma \in (0, 2\beta)$, then $(I - \gamma B)$ is averaged
  - $(I + \gamma A)^{-1}$ is $\frac{1}{2}$-averaged for any $\gamma > 0$

- **conclusion:** $T_{\text{FBS}}$ is averaged, thus

$$x^{k+1} \leftarrow T_{\text{FBS}}(x^k)$$

converges

\(^2\langle Bx - By, x - y \rangle \geq \beta \|Bx - By\|^2, \quad \forall x, y\)
The “big three” two-operator splitting schemes

- **forward-backward** (Mercier’79) for
  
  (maximally monotone) + (cocoercive)

- **Douglas-Rachford** (Lion-Mercier’79) for
  
  (maximally monotone) + (maximally monotone)

- **forward-backward-forward** (Tseng’00) for
  
  (maximally monotone) + (Lipschitz & monotone)

- all the schemes are built from **forward operators** and **backward operators**
Three-operator splitting
A three-operator splitting scheme

**motivation:**

- Nearly all existing splitting schemes reduce to one of the big three:
  - forward-backward (1970s)
  - Douglas-Rachford (1970s)
  - forward-backward-forward (2000)

- Given an $n$-operator problem? Reduce it to a bigger 2-operator splitting problem via extra variables

- Benefits of a native multi-operator splitting scheme
  - save extra variables, potential savings in memory and cpu time
  - fewer tricks, increased flexibility
  - improve theoretical understanding to operator splitting
A three-operator splitting scheme

- **require:** $A, B$ monotone, $C$ monotone and single-valued

- resolvents $J_A := (I + A)^{-1}$ and $J_B := (I + B)^{-1}$

- Davis and Yin’15:

  $$T_{DY} := I - J_{\gamma B} + J_{\gamma A} \circ (2J_{\gamma B} - I - \gamma C \circ J_{\gamma B})$$

  Computing $T_{DY}(z)$ will evaluate $J_{\gamma A}$, $J_{\gamma B}$, and $C$ only once each

- **encodes a fixed point:**

  $$0 \in Ax + Bx + Cx \iff z = T_{DY}(z), \quad x = J_{\gamma B}z$$
- Let $C$ be $\beta$-cocoercive, choose $\gamma \in (0, 2\beta)$

- Implement $z^{k+1} \leftarrow T_{DY}(z^k)$ as:
  1. $x_B^k \leftarrow J_{\gamma_B}(z^k)$
  2. $z^{k+1} \leftarrow z^k + (J_{\gamma_A}(2x_B^k - z^k - \gamma C x_B^k) - x_B^k)$

- $J_{\gamma_A}, J_{\gamma_B},$ and $\gamma C$ are evaluated once at each iteration
Convergence guarantee

- Apply $z^{k+1} \leftarrow T_{DY} z^k$ to
  
  $$0 \in A(x) + B(x) + C(x)$$

  where $C$ is $\beta$-cocoercive

- **convergence:** if a solution exists, then
  - any fixed point $z^* = T_{DY}(z^*)$ gives a solution $x^* = J_{\gamma B}(z^*)$
  - $z^k \rightarrow z^*$ weakly, if parameter $0 < \gamma < 2\beta$

- **speed:**
  - $1/\sqrt{k}$ non-ergodic and $1/k$ ergodic
  - improves to $1/k^2$ with strongly monotone $B$ or $C$
  - improves to $C^k$, $C < 1$, under stronger conditions
Three-operator applications
Three-operator direct applications

- Nonnegative matrix completion: recover $X^* \geq 0$ from

  $$y = A(X^*) + w, \quad w \sim N(0, \sigma)$$

  Problem formulation:

  $$\min_{X \in \mathbb{R}^{d \times m}} \|y - A(X)\|_F^2 + \mu \|X\|_* + \nu(X)$$

- Smooth optimization with linear and box constraints

  $$\min_{x \in \mathcal{H}} f(x) \quad \text{subject to: } Lx = b; \ 0 \leq x \leq b.$$  

  Special cases: kernelized SVMs and quadratic programs
Three-set split feasibility problem

- Find \( x \in \mathcal{H} \) such that
  \[
  x \in C_1 \cap C_2 \quad \text{and} \quad Lx \in C_3,
  \]

- Applications: **nonnegative semi-definite programs** and **conic programs** through homogeneous self-dual embedding.

- Problem reformulation:
  \[
  \min_x \langle C_1(x) + \langle C_2(x) + \frac{1}{2} \text{dist}_{\|\cdot\|_2}^2(Lx, C_3)^2
  \]
3-block ADMM

- problem:

\[
\begin{align*}
\text{minimize} & \quad f_1(x_1) + f_2(x_2) + f_3(x_3) \\
\text{subject to} & \quad L_1 x_1 + L_2 x_3 + L_3 x_3 = b,
\end{align*}
\]

Directly extended ADMM may fail to converge (Chen-He-Ye-Yuan’12)

- If \( f_1 \) is strongly convex, then apply Davis-Yin (to dual problem) gives:
  1. \( x_1^{k+1} = \arg \min_{x_1} L(x_1, x_2^k, x_3^k; w^k) \), Lagrangian
  2. \( x_2^{k+1} \in \arg \min_{x_2} L_\gamma(x_1^{k+1}, x_2, x_3^k; w^k) \), augmented Lagrangian
  3. \( x_3^{k+1} \in \arg \min_{x_3} L_\gamma(x_1^{k+1}, x_2^{k+1}, x_3; w^k) \), augmented Lagrangian
  4. \( w^{k+1} = w^k + \gamma(L_1 x_1^{k+1} + L_2 x_2^{k+1} + L_3 x_3^{k+1} - b) \)
not covered:

- Combining with other techniques to obtain algorithms for more applications
- Parallel algorithms for big problems
- Distributed computation

open questions:

- Remove the cocoercive assumption
- Four or more operator splitting

acknowledgements: NSF and ONR

report and slides at: http://www.math.ucla.edu/~wotaoyin/