Trajectory of Alternating Direction Method of Multipliers and Adaptive Acceleration

Clarice Poon (University of Bath)
Jingwei Liang (University of Cambridge)
Question: How should one accelerate the convergence of ADMM?

Constrained and composite optimisation problem:

\[ \min_{x \in \mathbb{R}^n, y \in \mathbb{R}^m} R(x) + J(y) \text{ such that } Ax + By = b \]  

(\mathcal{P})

under basic assumptions

- \( R, J \) are proper, convex, lower semi-continuous functions.
- \( A : \mathbb{R}^n \to \mathbb{R}^p \) and \( B : \mathbb{R}^m \to \mathbb{R}^p \) are injective linear operators.
- \( \text{ri} (\text{dom}(R) \cap \text{dom}(J)) \neq \emptyset \) and the set of minimizers is non-empty.
Given a fixed point sequence $z_{k+1} = \mathcal{F}(z_k)$, accelerate by

$$\tilde{z}_k = z_k + a_k(z_k - z_{k-1}), \quad a_k > 0,$$

$$z_{k+1} = \mathcal{F}(\tilde{z}_k).$$
The use of inertial

Given a fixed point sequence $z_{k+1} = \mathcal{F}(z_k)$, accelerate by

$$\bar{z}_k = z_k + a_k(z_k - z_{k-1}), \quad a_k > 0,$$

$$z_{k+1} = \mathcal{F}(\bar{z}_k).$$

Inertial is well-studied for algorithms such as gradient descent and Forward-Backward.

Improves the objective convergence rate from $O(k^{-1})$ to $O(k^{-2})$.

[Heavy-Ball/Nesterov accelerated gradient/FISTA]
The use of inertial

Given a fixed point sequence $z_{k+1} = \mathcal{F}(z_k)$, accelerate by

$$\tilde{z}_k = z_k + a_k(z_k - z_{k-1}), \quad a_k > 0,$$

$$z_{k+1} = \mathcal{F}(\tilde{z}_k).$$

Inertial is well-studied for algorithms such as gradient descent and Forward-Backward.

Improves the objective convergence rate from $O(k^{-1})$ to $O(k^{-2})$.

[Heavy-Ball/Nesterov accelerated gradient/FISTA]

Most works on inertial-ADMM impose extra assumptions (e.g. smoothness, uniform convexity).
The use of inertial

Given a fixed point sequence
\[ z_{k+1} = F(z_k), \]
accelerate by
\[ \tilde{z}_k = z_k + a_k(z_k - z_{k-1}), \quad a_k > 0, \]
\[ z_{k+1} = F(\tilde{z}_k). \]

Inertial is well-studied for algorithms such as gradient descent and Forward-Backward.

Improves the objective convergence rate from \( \mathcal{O}(k^{-1}) \) to \( \mathcal{O}(k^{-2}) \).

[Heavy-Ball/Nesterov accelerated gradient/FISTA]

Most works on inertial-ADMM impose extra assumptions (e.g. smoothness, uniform convexity).

The performance of inertial-ADMM in general is less clear.
Our contributions

1. We study the local trajectory of a sequence generated by ADMM under the framework of partial smoothness.
Our contributions

1. We study the local trajectory of a sequence generated by ADMM under the framework of partial smoothness.

Based on this trajectory analysis:

2. We obtain insight into when inertial will work and fail.
Our contributions

1. We study the local trajectory of a sequence generated by ADMM under the framework of partial smoothness.

Based on this trajectory analysis:

2. We obtain insight into when inertial will work and fail.

3. We develop an acceleration scheme with local acceleration rates.
The ADMM iterates

**Augmented Lagrangian:** For $\gamma > 0$ and Lagrangian multiplier $\psi \in \mathbb{R}^p$

$$\mathcal{L}(x, y, \psi) \overset{\text{def.}}{=} R(x) + J(y) + \langle \psi, Ax + By - b \rangle + \frac{\gamma}{2} \|Ax + By - b\|^2.$$

The ADMM iterations:

$$x_k = \text{argmin}_{x \in \mathbb{R}^n} R(x) + \frac{\gamma}{2} \|Ax + By_{k-1} - b + \frac{1}{\gamma} \psi_{k-1}\|^2,$$

$$y_k = \text{argmin}_{y \in \mathbb{R}^m} J(y) + \frac{\gamma}{2} \|Ax_k + By - b + \frac{1}{\gamma} \psi_{k-1}\|^2,$$

$$\psi_k = \psi_{k-1} + \gamma(Ax_k + By_k - b).$$
The ADMM iterates

Augmented Lagrangian: For $\gamma > 0$ and Lagrangian multiplier $\psi \in \mathbb{R}^p$

$$
\mathcal{L}(x, y, \psi) \overset{\text{def.}}{=} R(x) + J(y) + \langle \psi, Ax + By - b \rangle + \frac{\gamma}{2} \|Ax + By - b\|^2.
$$

The ADMM iterations:

$$
x_k = \arg\min_{x \in \mathbb{R}^n} R(x) + \frac{\gamma}{2} \|Ax + By_{k-1} - b + \frac{1}{\gamma} \psi_{k-1}\|^2,
$$

$$
y_k = \arg\min_{y \in \mathbb{R}^m} J(y) + \frac{\gamma}{2} \|Ax_k + By - b + \frac{1}{\gamma} \psi_{k-1}\|^2,
$$

$$
\psi_k = \psi_{k-1} + \gamma (Ax_k + By_k - b).
$$

Define $z_k \overset{\text{def.}}{=} \psi_{k-1} + \gamma Ax_k$. 

The ADMM iterates

**Augmented Lagrangian:** For $\gamma > 0$ and Lagrangian multiplier $\psi \in \mathbb{R}^p$

$$
\mathcal{L}(x, y, \psi) \overset{\text{def.}}{=} R(x) + J(y) + \langle \psi, Ax + By - b \rangle + \frac{\gamma}{2} \|Ax + By - b\|^2.
$$

The ADMM iterations:

$$
x_k = \text{argmin}_{x \in \mathbb{R}^n} R(x) + \frac{\gamma}{2} \|Ax - \frac{1}{\gamma}(z_{k-1} - 2\psi_{k-1})\|^2,
$$

$$
z_k = \psi_{k-1} + \gamma Ax_k,
$$

$$
y_k = \text{argmin}_{y \in \mathbb{R}^m} J(y) + \frac{\gamma}{2} \|By + \frac{1}{\gamma}(z_k - \gamma b)\|^2,
$$

$$
\psi_k = z_k + \gamma (By_k - b).
$$

Then, $z_k = \mathcal{F}(z_{k-1})$ for some fixed point operator $\mathcal{F} \dagger$.

\dagger Due to the equivalence between ADMM and Douglas-Rachford splitting [Gabay '83].
The ADMM iterates

**Augmented Lagrangian:** For $\gamma > 0$ and Lagrangian multiplier $\psi \in \mathbb{R}^p$

$$\mathcal{L}(x, y, \psi) \overset{\text{def.}}{=} R(x) + J(y) + \langle \psi, Ax + By - b \rangle + \frac{\gamma}{2} \|Ax + By - b\|^2.$$

The ADMM iterations:

$$x_k = \arg\min_{x \in \mathbb{R}^n} R(x) + \frac{\gamma}{2} \|Ax - \frac{1}{\gamma}(z_{k-1} - 2\psi_{k-1})\|^2,$$

$$z_k = \psi_{k-1} + \gamma Ax_k,$$

$$y_k = \arg\min_{y \in \mathbb{R}^m} J(y) + \frac{\gamma}{2} \|By + \frac{1}{\gamma}(z_k - \gamma b)\|^2,$$

$$\psi_k = z_k + \gamma (By_k - b).$$

We will analyse the behaviour of $\{z_k\}_k$. 
Partial smoothness [Lewis ’05]

\( R \) is partly smooth at \( x \) relative to a set \( \mathcal{M} \ni x \) if \( \partial R(x) \neq \emptyset \) and

**Smoothness:**
\( \mathcal{M} \) is a \( C^2 \)-manifold, \( R|_{\mathcal{M}} \) is \( C^2 \) near \( x \).

**Sharpness:**
Tangent space \( T_{\mathcal{M}}(x) \) is \( \text{par} \left( \partial R(x) \right)^\perp \).

**Continuity:**
\( \partial R \) is continuous along \( \mathcal{M} \) near \( x \).

\( \text{par}(C) \): sub-space parallel to \( C \), where \( C \) is a non-empty convex set.

\( \text{PSF}_x(\mathcal{M}_x) \): function that is partly smooth at \( x \) relative to \( \mathcal{M}_x \).

**Examples:** \( \ell_1, \ell_{1,2}, \ell_\infty \)-norm, nuclear norm, total variation.
Partial smoothness

If \( R \in \text{PSF}_{x^*}(\mathcal{M}^R_{x^*}) \) and \( J \in \text{PSF}_{y^*}(\mathcal{M}^J_{y^*}) \), then under non-degeneracy conditions around \( x^* \) and \( y^* \):

**Manifold identification and local linearisation** [Liang, Fadili & Peyré ’16]:

There exists \( K \in \mathbb{N} \) and a matrix \( M_{\text{ADMM}} \) such that for all \( k \geq K \),

- \( x_k \in \mathcal{M}^R_{x^*} \) and \( y_k \in \mathcal{M}^J_{y^*} \).

- \( z_k - z^* = M_{\text{ADMM}}(z_{k-1} - z^*) + o(\|z_{k-1} - z^*\|) \).
Partial smoothness

If $R \in \text{PSF}_{x^*}(\mathcal{M}_{x^*}^R)$ and $J \in \text{PSF}_{y^*}(\mathcal{M}_{y^*}^J)$, then under non-degeneracy conditions around $x^*$ and $y^*$:

**Manifold identification and local linearisation** [Liang, Fadili & Peyré ’16]:
There exists $K \in \mathbb{N}$ and a matrix $M_{ADMM}$ such that for all $k \geq K$,

- $x_k \in \mathcal{M}_{x^*}^R$ and $y_k \in \mathcal{M}_{y^*}^J$.
- $z_k - z^* = M_{ADMM}(z_{k-1} - z^*) + o(\|z_{k-1} - z^*\|)$.

The behaviour of $z_k$ is eventually regular.
Let $v_k \overset{\text{def.}}{=} z_k - z_{k-1}$ and $\theta_k = \angle (v_k, v_{k-1})$.

### Two non-smooth terms

$R$ and $J$ are locally polyhedral around $x^*$ and $y^*$.

### Spiral trajectory:

$$
\cos(\theta_k) = \cos(\alpha) + O(\eta^{2k})
$$

with $\eta < 1$, $\alpha > 0$.

$M_{\text{ADMM}}$ has complex eigenvalues.

### At least one smooth term

$A$ is an invertible square matrix and $R$ is locally $C^2$ around $x^*$.

### Straight line trajectory:

$$
\cos(\theta_k) \to 1 \text{ when } \gamma > \| (A^\top A)^{-\frac{1}{2}} \nabla^2 R(x^*)(A^\top A)^{-\frac{1}{2}} \|.
$$

$M_{\text{ADMM}}$ has all real eigenvalues.
One inertial-ADMM iteration:

\[ x_k = \arg\min_{x \in \mathbb{R}^n} R(x) + \frac{\gamma}{2} \| Ax - \frac{1}{\gamma} (\bar{z}_{k-1} - 2\psi_{k-1}) \|^2, \]

\[ z_k = \psi_{k-1} + \gamma Ax_k, \]

\[ \bar{z}_k = z_k + a_k (z_k - z_{k-1}), \]

\[ y_k = \arg\min_{y \in \mathbb{R}^m} J(y) + \frac{\gamma}{2} \| By + \frac{1}{\gamma} (\bar{z}_k - \gamma b) \|^2, \]

\[ \psi_k = \bar{z}_k + \gamma (By_k - b). \]

**Intuition:** inertial-ADMM accelerates if \( z_k \) is moving along a straight path...
Failure of inertial-ADMM

Find $x \in T_1 \cap T_2$. Solve using ADMM

$$\min_{x,y} \nu_{T_1}(x) + \nu_{T_2}(y) \quad \text{such that} \quad x - y = 0.$$ 

Consider $z_k \overset{\text{def.}}{=} \psi_{k-1} + \gamma x_k$. **Standard ADMM:**

![Diagram showing iterative process and convergence](image-url)
Find \( x \in T_1 \cap T_2 \). Solve using ADMM

\[
\min_{x,y} \nu_{T_1}(x) + \nu_{T_2}(y) \quad \text{such that} \quad x - y = 0.
\]

Consider \( z_k \overset{\text{def.}}{=} \psi_{k-1} + \gamma x_k \). Inertial-ADMM with \( a = 0.25 \):
Failure of inertial-ADMM

LASSO example:

$$\min_{x,y \in \mathbb{R}^n} \mu \|x\|_1 + \frac{1}{2} \|Ky - f\|_2^2$$

such that \( x - y = 0 \).

$$\gamma = \frac{\|K\|^2}{10}$$

$$\gamma = \|K\|^2 + 0.1$$

Eventual trajectory:
- Straight line when \( \gamma > \|K\|^2 \)
- \( M_{\text{ADMM}} \) may have complex leading eigenvalue if \( \gamma \leq \|K\|^2 \).
Idea: Given past points \( \{ z_{k-j} \}_{j=0}^{q+1} \), define \( \{ v_{k-j} \}_{j=0}^{q} \) as:

\[
v_{k-j} \overset{\text{def.}}{=} z_{k-j} - z_{k-j-1}
\]

- Fit the past directions \( v_{k-1}, \ldots, v_{k-q} \) to the latest direction \( v_k \):

\[
c^k \overset{\text{def.}}{=} \arg\min_{c \in \mathbb{R}^q} \| \sum_{j=1}^{q} c_j v_{k-j} - v_k \|^2.
\]
Adaptive acceleration for ADMM ($A^3$DMM)

**Idea:** Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, define $\{v_{k-j} \overset{\text{def.}}{=} z_{k-j} - z_{k-1-j}\}_{j=0}^{q}$.

- Fit the past directions $v_{k-1}, \ldots, v_{k-q}$ to the latest direction $v_k$:
  
  $$c^k \overset{\text{def.}}{=} \arg\min_{c \in \mathbb{R}^q} \| \sum_{j=1}^{q} c_j v_{k-j} - v_k \|^2.$$  

- Let $\bar{z}_{k,1} \overset{\text{def.}}{=} z_k + \sum_{j=1}^{q} c_j v_{k-j+1}$.
Adaptive acceleration for ADMM ($A^3$DMM)

**Idea:** Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, define $\{v_{k-j} \overset{\text{def.}}{=} z_{k-j} - z_{k-j-1}\}_{j=0}^{q}$. 

- Fit the past directions $v_{k-1}, \ldots, v_{k-q}$ to the latest direction $v_k$:
  \[ c^k \overset{\text{def.}}{=} \arg\min_{c \in \mathbb{R}^q} \| \sum_{j=1}^{q} c_j v_{k-j} - v_k \|^2. \]

- Let $\bar{z}_{k,1} \overset{\text{def.}}{=} z_k + \sum_{j=1}^{q} c^k_j v_{k-j+1}$.

Repeat on $\{z_{k-j}\}_{j=0}^{q} \cup \{\bar{z}_{k,1}\}$ and so on.
Adaptive acceleration for ADMM ($A^3DMM$)

**Idea:** Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, define $\{v_{k-j} \text{ def.} = z_{k-j} - z_{k-j-1}\}_{j=0}^{q}$.

- Fit the past directions $v_{k-1}, \ldots, v_{k-q}$ to the latest direction $v_k$:
  $$c^k \text{ def.} = \arg\min_{c \in \mathbb{R}^q} \| \sum_{j=1}^{q} c_j v_{k-j} - v_k \|^2.$$ 
- Let $\bar{z}_{k,1} \text{ def.} = z_k + \sum_{j=1}^{q} c_j^k v_{k-j+1}$.

Repeat on $\{z_{k-j}\}_{j=0}^{q} \cup \{\bar{z}_{k,1}\}$ and so on.

The *s-step extrapolation* is $\bar{z}_{k,s} = z_k + \mathcal{E}_{s,q,k}$, where $\mathcal{E}_{s,q,k} = \sum_{j=1}^{q} \hat{c}_j v_{k-j+1}$ and

$$\hat{c} \text{ def.} = \left( \sum_{j=1}^{s} H(c^k)^j \right)_{(;1)} \text{ with } H(c^k) \text{ def.} = \begin{bmatrix} c^k & \frac{\text{Id}_{q-1}}{0_{1,q-1}} \end{bmatrix}.$$
Initial: Let $s \geq 1$, $q \geq 1$. Let $\bar{z}_0 = z_0 \in \mathbb{R}^p$ and $V_0 = O_{p \times (q+1)}$.

Repeat: For $k \geq 1$

- $y_k = \arg\min_{y \in \mathbb{R}^m} J(y) + \frac{\gamma}{2} \| By + \frac{1}{\gamma} (\bar{z}_{k-1} - \gamma b) \|^2$,
- $\psi_k = \bar{z}_{k-1} + \gamma (By_k - b)$,
- $x_k = \arg\min_{x \in \mathbb{R}^n} R(x) + \frac{\gamma}{2} \| Ax - \frac{1}{\gamma} (\bar{z}_{k-1} - 2\psi_k) \|^2$,
- $z_k = \psi_k + \gamma Ax_k$,
- $v_k = z_k - z_{k-1}$ and $V_k = [v_k, V_k(:, 1:q)]$.

If $\text{mod}(k, q + 2) = 0$: Compute coefficients $c^k$ and let $C_k \overset{\text{def.}}{=} H(c^k)$

- If $\rho(C_k) < 1$: $\bar{z}_k = z_k + a_k E_{s,q,k}$; else: $\bar{z}_k = z_k$.

If $\text{mod}(k, q + 2) \neq 0$: $\bar{z}_k = z_k$.
Global convergence is guaranteed for appropriate choice of $a_k$.

Local acceleration depends on $\varepsilon_k \overset{\text{def.}}{=} \min_c \| V_{k-1}c - v_k \|$.

- If $M_{\text{ADMM}}$ is diagonalisable, then $\varepsilon_k = \mathcal{O}(|\lambda_{q+1}|^k)$ where $\lambda_{q+1}$ is the $(q + 1)^{th}$ largest eigenvalue.
- Guaranteed local acceleration for $q = 2$ if $R$ and $J$ are polyhedral.

Related to vector extrapolation techniques from the 1960’s.

[Aitken ’27, Wynn ’62, Andersen ’65...]
Remarks

Implementation:

- Typically set $q \leq 10$.
- Extra memory cost of $p \times (q + 1)$ (storing $V_k$).
- Extra computation cost of $q^2 p$ every $(q + 2)$ iterations.
- One could also extrapolate $\{x_k, y_k\}$ simultaneously. But this would require extra storage of past directions.
Experiment: 2 non-smooth terms

Basis pursuit type problem with $\Omega \overset{\text{def.}}{=} \{ x \in \mathbb{R}^n : Kx = f \}$:

$$\min_{x, y \in \mathbb{R}^n} R(x) + \iota_\Omega(y) \quad \text{such that} \quad x - y = 0.$$
Inertial ADMM is slower than ADMM as eventual trajectory is a spiral.
The LASSO problem

$$\min_{x, y \in \mathbb{R}^n} R(x) + \frac{1}{2} \| Ky - f \|^2 \text{ such that } x - y = 0.$$
Inertial ADMM does accelerate, but $A^3$DMM is significantly faster.
Experiment: Total variation based image inpainting

Let $\Omega \overset{\text{def.}}{=} \{ x \in \mathbb{R}^{n \times n} : P_D(x) = f \}$, $P_D$ randomly sets 50% pixels to zero and consider

$$\min_{x \in \mathbb{R}^{n \times n}} \| y \|_1 + \iota_\Omega(x) \quad \text{such that} \quad \nabla x - y = 0.$$ 

- Both functions are polyhedral, trajectory is a spiral.
- Inertial ADMM is slower than ADMM.
Experiment: Total variation based image inpainting

Original image

ADMM, PSNR = 26.5448

Inertial ADMM, PSNR = 26.1096

Corrupted image

$A^3_{\text{DMM}} s = 100, \text{PSNR} = 27.0402$

$A^3_{\text{DMM}} s = +\infty, \text{PSNR} = 27.0402$
Summary of contributions

Trajectory of ADMM  For sequence $\{z_k\}_{k \in \mathbb{N}}$

- When both $R$ and $J$ are locally polyhedral around the fixed point, $\{z_k\}_{k \in \mathbb{N}}$ eventually moves along a **spiral**.
- When at least one of $R$ or $J$ is smooth, the trajectory of $\{z_k\}_{k \in \mathbb{N}}$ depends on $\gamma$ and can be either a spiral or a **straight line**.
Summary of contributions

Trajectory of ADMM  For sequence \( \{z_k\}_{k \in \mathbb{N}} \)
- When both \( R \) and \( J \) are locally polyhedral around the fixed point, \( \{z_k\}_{k \in \mathbb{N}} \) eventually moves along a **spiral**.
- When at least one of \( R \) or \( J \) is smooth, the trajectory of \( \{z_k\}_{k \in \mathbb{N}} \) depends on \( \gamma \) and can be either a spiral or a **straight line**.

An adaptive acceleration for ADMM
- The different trajectory behaviour of ADMM can lead to the **failure** of the inertial technique.
- We propose an acceleration strategy based on the idea of following the sequence trajectory.
Summary of contributions

Trajectory of ADMM  For sequence $\{z_k\}_{k \in \mathbb{N}}$

- When both $R$ and $J$ are locally polyhedral around the fixed point, $\{z_k\}_{k \in \mathbb{N}}$ eventually moves along a **spiral**.
- When at least one of $R$ or $J$ is smooth, the trajectory of $\{z_k\}_{k \in \mathbb{N}}$ depends on $\gamma$ and can be either a spiral or a **straight line**.

An adaptive acceleration for ADMM

- The different trajectory behaviour of ADMM can lead to the **failure** of the inertial technique.
- We propose an acceleration strategy based on the idea of following the sequence trajectory.

Poster: East Exhibition Hall B+C #115!