Trajectory of Alternating Direction Method of Multipliers and Adaptive Acceleration

Clarice Poon (University of Bath)
Jingwei Liang (University of Cambridge)

Alternating Direction Method of Multipliers (ADMM)

Constrained and composite optimisation problem:

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}^m} R(x) + J(y) \quad \text{such that} \quad Ax + By = b \tag{$\mathcal{P}$}$$

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.
- $\mathrm{ri}\,(\mathrm{dom}(R) \cap \mathrm{dom}(J)) \neq \emptyset$ and the set of minimizers is non-empty.

Constrained and composite optimisation problem:

$$\min_{x \in \mathbb{R}^n, y \in \mathbb{R}^m} R(x) + J(y) \quad \text{such that} \quad Ax + By = b \tag{$\mathcal{P}$}$$

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.
- $\mathrm{ri}\,(\mathrm{dom}(R) \cap \mathrm{dom}(J)) \neq \emptyset$ and the set of minimizers is non-empty.

Question: *How should one accelerate the convergence of ADMM?*

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

$$\begin{aligned} \bar{z}_k &= z_k + \alpha_k(z_k - z_{k-1}), \quad \alpha_k > 0, \\ z_{k+1} &= \mathcal{F}(\bar{z}_k) \end{aligned}$$

The use of inertial

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

$$\begin{aligned} \bar{z}_k &= z_k + \alpha_k(z_k - z_{k-1}), \quad \alpha_k > 0, \\ z_{k+1} &= \mathcal{F}(\bar{z}_k) \end{aligned}$$

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]

The use of inertial

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

$$\begin{aligned} \bar{z}_k &= z_k + \alpha_k(z_k - z_{k-1}), \quad \alpha_k > 0, \\ z_{k+1} &= \mathcal{F}(\bar{z}_k) \end{aligned}$$

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 use of inertial

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

$$\begin{aligned} \bar{z}_k &= z_k + \alpha_k(z_k - z_{k-1}), \quad \alpha_k > 0, \\ z_{k+1} &= \mathcal{F}(\bar{z}_k) \end{aligned}$$

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.

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 $\psi \in \mathbb{R}^p$

$$\mathcal{L}(x, y, \psi) \stackrel{\text{def.}}{=} R(x) + J(y) + \langle \psi,\ Ax + By - b \rangle + \frac{\gamma}{2} \|Ax + By - b\|_2^2.$$

The ADMM iterations:

$$\begin{aligned}
x_k &= \operatorname{argmin}_{x \in \mathbb{R}^n} R(x) + \frac{\gamma}{2} \|Ax + By_{k-1} - b + \tfrac{1}{\gamma} \psi_{k-1}\|^2, \\
y_k &= \operatorname{argmin}_{y \in \mathbb{R}^m} J(y) + \frac{\gamma}{2} \|Ax_k + By - b + \tfrac{1}{\gamma} \psi_{k-1}\|^2, \\
\psi_k &= \psi_{k-1} + \gamma(Ax_k + By_k - b).
\end{aligned}$$

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

$$\mathcal{L}(x, y, \psi) \stackrel{\text{def.}}{=} R(x) + J(y) + \langle \psi,\ Ax + By - b \rangle + \frac{\gamma}{2}\|Ax + By - b\|_2^2.$$

The ADMM iterations:

$$\begin{aligned}
x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \frac{\gamma}{2}\|Ax + By_{k-1} - b + \frac{1}{\gamma}\psi_{k-1}\|^2, \\
y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \frac{\gamma}{2}\|Ax_k + By - b + \frac{1}{\gamma}\psi_{k-1}\|^2, \\
\psi_k &= \psi_{k-1} + \gamma(Ax_k + By_k - b).
\end{aligned}$$

Define $z_k \stackrel{\text{def.}}{=} \psi_{k-1} + \gamma A x_k$.

The ADMM iterates

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

$$\mathcal{L}(x, y, \psi) \stackrel{\text{def.}}{=} R(x) + J(y) + \langle \psi,\, Ax + By - b \rangle + \frac{\gamma}{2} \|Ax + By - b\|_2^2.$$

The ADMM iterations:

$$\begin{aligned} x_k &= \mathrm{argmin}_{x \in \mathbb{R}^n}\, R(x) + \frac{\gamma}{2} \|Ax - \frac{1}{\gamma}(z_{k-1} - 2\psi_{k-1})\|^2, \\ z_k &= \psi_{k-1} + \gamma A x_k, \\ y_k &= \mathrm{argmin}_{y \in \mathbb{R}^m}\, J(y) + \frac{\gamma}{2} \|By + \frac{1}{\gamma}(z_k - \gamma b)\|^2, \\ \psi_k &= z_k + \gamma(B y_k - b). \end{aligned}$$

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

† Due to the equivalence between ADMM and Douglas-Rachford splitting [Gabay '83].

The ADMM iterates

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

$$\mathcal{L}(x, y, \psi) \stackrel{\text{def.}}{=} R(x) + J(y) + \langle \psi,\, Ax + By - b \rangle + \frac{\gamma}{2}\|Ax + By - b\|_2^2.$$

The ADMM iterations:

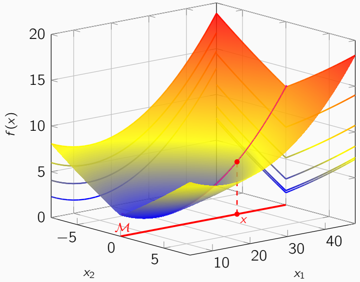
$$\begin{aligned}
x_k &= \operatorname{argmin}_{x \in \mathbb{R}^n} R(x) + \frac{\gamma}{2}\|Ax - \frac{1}{\gamma}(z_{k-1} - 2\psi_{k-1})\|^2, \\
z_k &= \psi_{k-1} + \gamma A x_k, \\
y_k &= \operatorname{argmin}_{y \in \mathbb{R}^m} J(y) + \frac{\gamma}{2}\|By + \frac{1}{\gamma}(z_k - \gamma b)\|^2, \\
\psi_k &= z_k + \gamma(By_k - b).
\end{aligned}$$

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 $\mathcal{T}_{\mathcal{M}}(x)$ is $\mathrm{par}\left(\partial R(x)\right)^{\perp}$.
Continuity:
∂R is continuous along $\mathcal{M}$ near x.



$\mathrm{par}(C)$: sub-space parallel to C, where C is a non-empty convex set.
$\mathrm{PSF}_x(\mathcal{M}_x)$: 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 $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}^R_{x^\star})$ and $J \in \mathrm{PSF}_{y^\star}(\mathcal{M}^J_{y^\star})$, then under **non-degeneracy conditions** around $x^\star$ and $y^\star$:

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

- $x_k \in \mathcal{M}^R_{x^\star}$ and $y_k \in \mathcal{M}^J_{y^\star}$
- $z_k - z^\star = M_{\text{ADMM}}(z_{k-1} - z^\star) + o(\|z_{k-1} - z^\star\|)$

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

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

- $x_k \in \mathcal{M}^R_{x^\star}$ and $y_k \in \mathcal{M}^J_{y^\star}$
- $z_k - z^\star = M_{\text{ADMM}}(z_{k-1} - z^\star) + o(\|z_{k-1} - z^\star\|)$

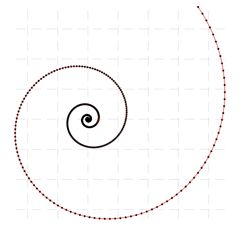
The behaviour of z_k is eventually **regular**.

Partial smoothness and sequence trajectory

Let $v_k \stackrel{\text{def.}}{=} z_k - z_{k-1}$ and $\theta_k = \angle(v_k, v_{k-1})$.

Two non-smooth terms

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



Spiral trajectory:

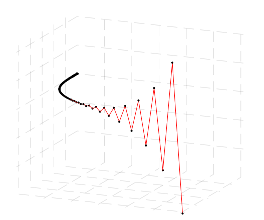
$$\cos(\theta_k) = \cos(\alpha) + \mathcal{O}(\eta^{2k})$$

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

M_{ADMM} has **complex** eigenvalues

At least one smooth term

A is a full rank square matrix and R is locally C^2 around $x^\star$.



Straight line trajectory:

$\cos(\theta_k) \to 1$ when

$$\gamma > \|(A^\top A)^{-\frac{1}{2}} \nabla^2 R(x^\star)(A^\top A)^{-\frac{1}{2}}\|.$$

M_{ADMM} has all **real** eigenvalues

One inertial-ADMM iteration:

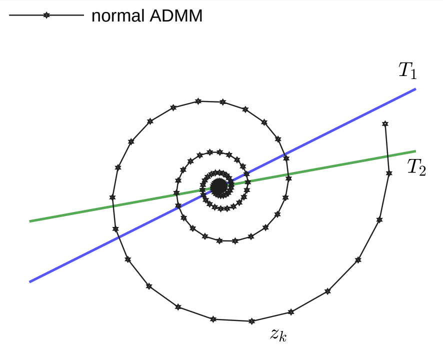
$$
\begin{aligned}
x_k &= \operatorname{argmin}_{x\in\mathbb{R}^n} R(x) + \tfrac{\gamma}{2}\|Ax - \tfrac{1}{\gamma}(\bar{z}_{k-1} - 2\psi_{k-1})\|^2, \\
z_k &= \psi_{k-1} + \gamma A x_k, \\
\bar{z}_k &= z_k + a_k(z_k - z_{k-1}), \\
y_k &= \operatorname{argmin}_{y\in\mathbb{R}^m} J(y) + \tfrac{\gamma}{2}\|By + \tfrac{1}{\gamma}(\bar{z}_k - \gamma b)\|^2, \\
\psi_k &= \bar{z}_k + \gamma(By_k - b).
\end{aligned}
$$

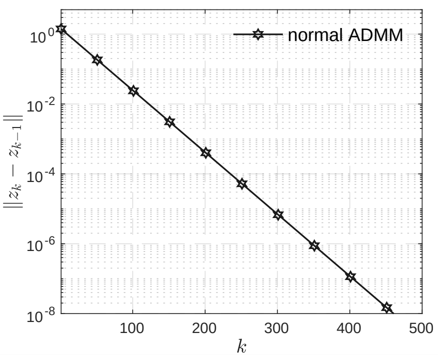
Intuition: inertial-ADMM accelerates if z_k is moving along a straight path...

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

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

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



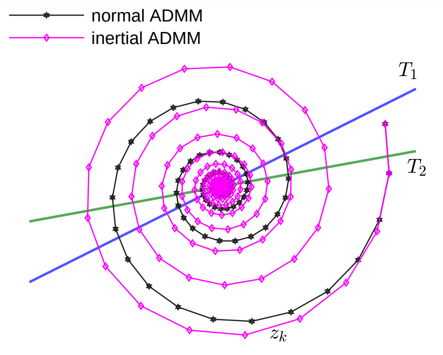


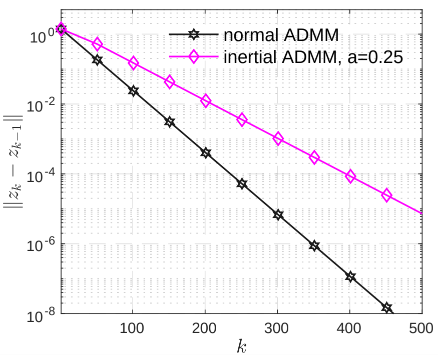
Failure of inertial-ADMM

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

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

Consider $z_k \stackrel{\text{def.}}{=} \psi_{k-1} + \gamma x_k$. **Inertial-ADMM with $a = 0.25$:**



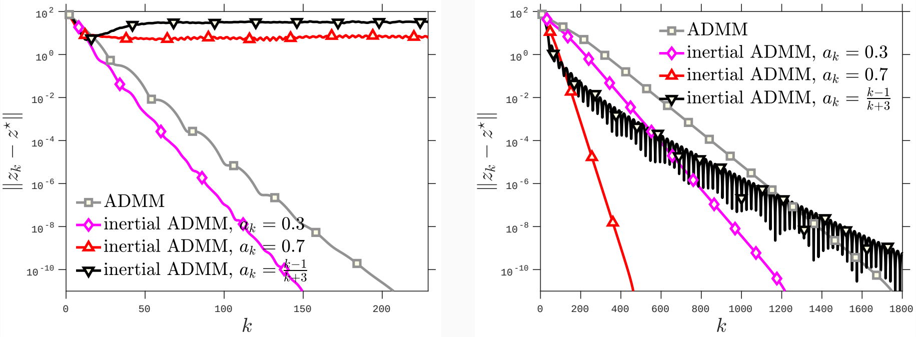


Failure of inertial-ADMM

LASSO example:

$$\min_{x,y\in\mathbb{R}^n} \mu\|x\|_1 + \tfrac{1}{2}\|Ky - f\|_2^2 \quad \text{such that} \quad x - y = 0.$$

$\gamma = \|K\|^2/10$ $\qquad\qquad$ $\gamma = \|K\|^2 + 0.1$



Eventual trajectory:

- Straight line when $\gamma > \|K\|^2$
- M_{ADMM} may have complex leading eigenvalue if $\gamma \leqslant \|K\|^2$.

Adaptive acceleration for ADMM (A^{3}DMM)

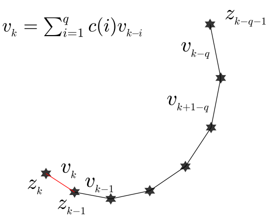
Goal: Given past points $\{z_{k-j}\}_{j=0}^{q}$, predict z_{k+1}.

Idea: Define $v_j \stackrel{\text{def.}}{=} z_j - z_{j-1}$,

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

$$c_k \stackrel{\text{def.}}{=} \operatorname{argmin}_{c \in \mathbb{R}^q} \|V_{k-1} c - v_k\|^2,$$

where $V_{k-1} = [v_{k-1}, \ldots, v_{k-q}] \in \mathbb{R}^{n \times q}$.



Goal: Given past points $\{z_{k-j}\}_{j=0}^{q}$, predict z_{k+1}.

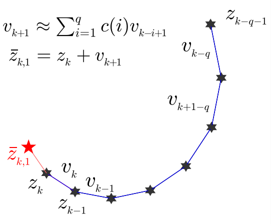
Idea: Define $v_j \stackrel{\text{def.}}{=} z_j - z_{j-1}$,

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

$$c_k \stackrel{\text{def.}}{=} \mathrm{argmin}_{c \in \mathbb{R}^q} \|V_{k-1}c - v_k\|^2,$$

where $V_{k-1} = [v_{k-1}, \ldots, v_{k-q}] \in \mathbb{R}^{n \times q}$.

S.2) If $V_k c_k \approx v_{k+1}$, then $\bar{z}_{k,1} \stackrel{\text{def.}}{=} z_k + V_k c_k \approx z_{k+1}$.



Adaptive acceleration for ADMM (A^{3}DMM)

Goal: Given past points $\{z_{k-j}\}_{j=0}^{q}$, predict z_{k+1}.

Idea: Define $v_j \overset{\text{def.}}{=} z_j - z_{j-1}$,

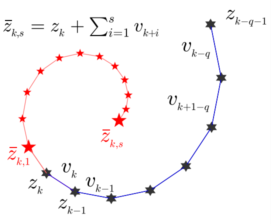
S.1) Fit the past directions $v_{k-1}, \ldots, v_{k-q}$ to the latest direction v_k:

$$c_k \overset{\text{def.}}{=} \operatorname{argmin}_{c \in \mathbb{R}^q} \|V_{k-1}c - v_k\|^2,$$

where $V_{k-1} = [v_{k-1}, \ldots, v_{k-q}] \in \mathbb{R}^{n \times q}$.

S.2) If $V_k c_k \approx v_{k+1}$, then $\bar{z}_{k,1} \overset{\text{def.}}{=} z_k + V_k c_k \approx z_{k+1}$.

Repeat s times to predict z_{k+s}.



Adaptive acceleration for ADMM (A^{3}DMM)

Goal: Given past points $\{z_{k-j}\}_{j=0}^{q}$, predict z_{k+1}.

Idea: Define $v_j \stackrel{\text{def.}}{=} z_j - z_{j-1}$,

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

$$c_k \stackrel{\text{def.}}{=} \operatorname{argmin}_{c \in \mathbb{R}^q} \|V_{k-1}c - v_k\|^2,$$

where $V_{k-1} = [v_{k-1}, \ldots, v_{k-q}] \in \mathbb{R}^{n \times q}$.

S.2) If $V_k c_k \approx v_{k+1}$, then $\bar{z}_{k,1} \stackrel{\text{def.}}{=} z_k + V_k c_k \approx z_{k+1}$.

Repeat s times to predict z_{k+s}.

Define: $H(c_k) \stackrel{\text{def.}}{=} \left[\begin{array}{c|c} c_k & \begin{array}{c} \mathrm{Id}_{q-1} \\ \hline 0_{1,q-1} \end{array} \end{array} \right]$ and $\mathcal{E}_{s,q,k} = V_k \left(\sum_{j=1}^{s} H(c_k)^j \right)_{(:,1)}$.

The ***s*-step extrapolation** is $\bar{z}_{k,s} = z_k + \mathcal{E}_{s,q,k}$.

Initial: Let $s \geqslant 1$, $q \geqslant 1$, $\bar{q} = q + 1$. Let $\bar{z}_0 = z_0 \in \mathbb{R}^p$ and $V_0 = 0_{p\times q}$.

Repeat: For $k \geqslant 1$

$$\begin{aligned}
y_k &= \mathrm{argmin}_{y\in\mathbb{R}^m}\, J(y) + \tfrac{\gamma}{2}\|By + \tfrac{1}{\gamma}\left(\bar{z}_{k-1} - \gamma b\right)\|^2 \\
\psi_k &= \bar{z}_{k-1} + \gamma(By_k - b) \\
x_k &= \mathrm{argmin}_{x\in\mathbb{R}^n}\, R(x) + \tfrac{\gamma}{2}\|Ax - \tfrac{1}{\gamma}\left(\bar{z}_{k-1} - 2\psi_k\right)\|^2 \\
z_k &= \psi_k + \gamma A x_k \\
v_k &= z_k - z_{k-1} \quad \text{and} \quad V_k = [v_k, V_k(:, 1:q-1)]
\end{aligned}$$

If $\mathrm{mod}(k, \bar{q}) = 0$: Compute coefficients c_k and let $C_k \overset{\text{def.}}{=} H(c_k)$

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

If $\mathrm{mod}(k, \bar{q}) \neq 0$: $\bar{z}_k = z_k$.

Remarks

Global convergence is guaranteed for appropriate choice of a_k.

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

- If M_{ADMM} is diagonalisable, then $\varepsilon_k = \mathcal{O}(|\lambda_{\bar{q}}|^k)$ where $\lambda_{\bar{q}}$ is the $\bar{q}^{th}$ 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

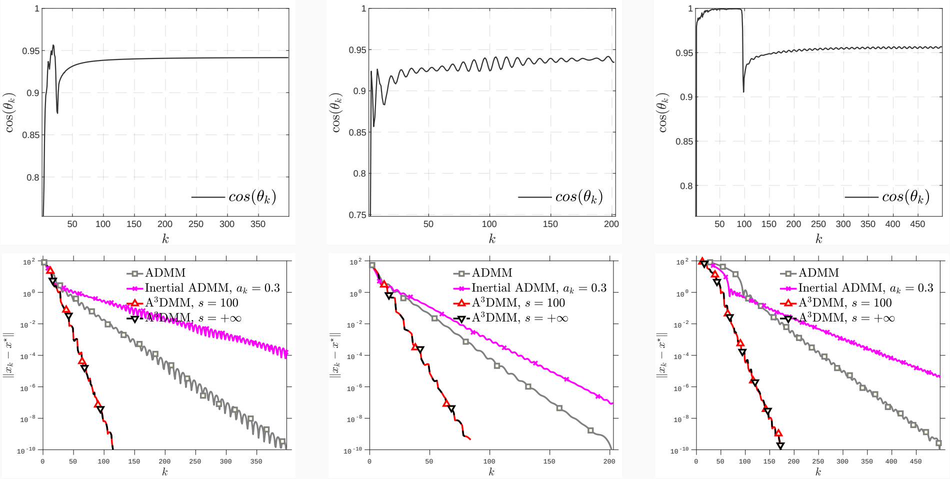
Implementation:

- Typically set $q \leqslant 10$.
- 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

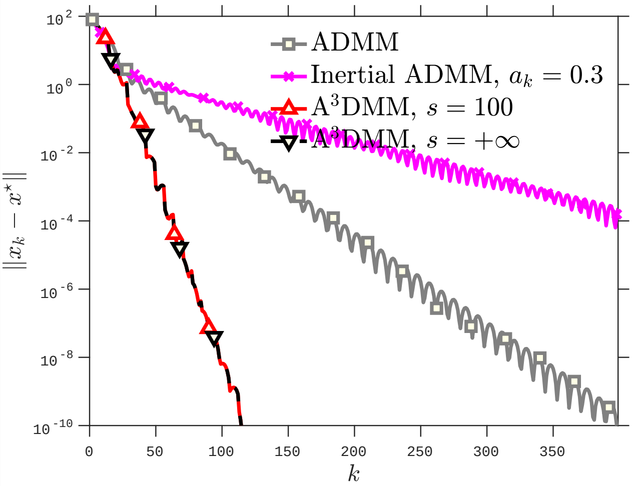
Basis pursuit type problem with $\Omega \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^n \;;\; Kx = f\}$:

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



ℓ_1-norm $\qquad$ $\ell_{1,2}$-norm $\qquad$ Nuclear norm

Experiment: 2 non-smooth terms

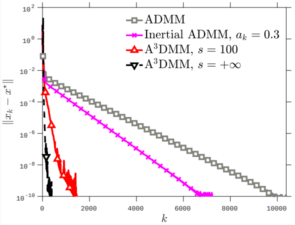


ℓ_1-norm

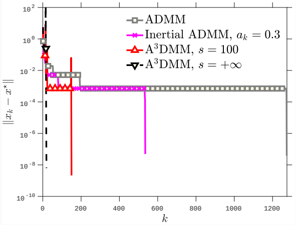
Inertial ADMM is **slower** than ADMM as eventual trajectory is a spiral.

Consider the LASSO problem

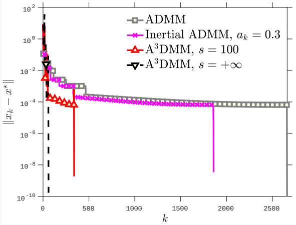
$$\min_{x,y\in\mathbb{R}^n} R(x) + \frac{1}{2}\|Ky - f\|^2 \quad \text{such that} \quad x - y = 0.$$



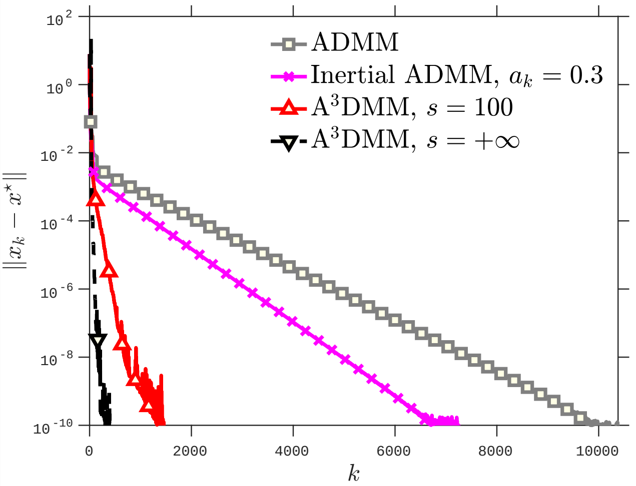
covtype



ijcnn1



phishing



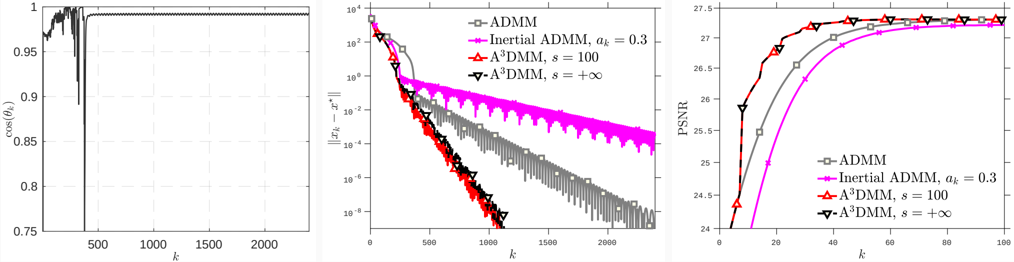
covtype

Inertial ADMM does accelerate, but A^3DMM is significantly faster.

Let $\Omega \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^{n\times n} \;;\; P_{\mathcal{D}}(x) = f\}$, $P_{\mathcal{D}}$ randomly sets 50% pixels to zero and consider

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

Angle $\{\theta_k\}$ $\qquad$ $\|x_k - x^\star\|$ $\qquad$ PSNR



- Both functions are polyhedral, trajectory is a spiral.
- Inertial ADMM is **slower** than ADMM.

Experiment: Total variation based image inpainting

Original image

ADMM, PSNR = 26.6935

Inertial ADMM, PSNR = 26.3203

Corrupted image

A^{3}DMM $s = 100$, PSNR = 27.1668

A^{3}DMM $s = +\infty$, PSNR = 27.1667

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 γ 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 γ 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 γ 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!