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:

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

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.

The use of inertial

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

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

The use of inertial

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

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

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

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

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

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

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.

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.

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:

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

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:

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

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:

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

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

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

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:

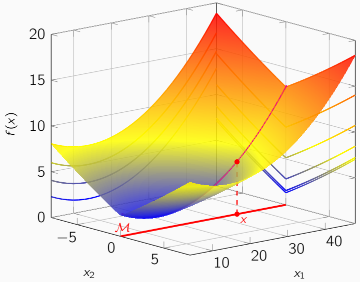
$\mathcal{M}$ is a C^2-manifold, $R|_{\mathcal{M}}$ is C^2 near x.

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.

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 $y_k \in \mathcal{M}^J_{y^\star}$.
- $z_k - z^\star = M_{\mathrm{ADMM}}(z_{k-1} - z^\star) + o(\|z_{k-1} - z^\star\|)$.

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 $y_k \in \mathcal{M}^J_{y^\star}$.
- $z_k - z^\star = M_{\mathrm{ADMM}}(z_{k-1} - z^\star) + o(\|z_{k-1} - z^\star\|)$.

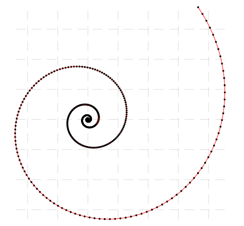
The behaviour of z_k is eventually **regular**.

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:

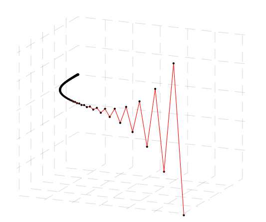
$$\cos(\theta_k) = \cos(\alpha) + \mathcal{O}(\eta^{2k})$$

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

M_{ADMM} has **complex** eigenvalues

At least one smooth term

A is an invertible square matrix and R is locally $\mathcal{C}^2$ 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

Partial smoothness and sequence trajectory

One inertial-ADMM iteration:

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

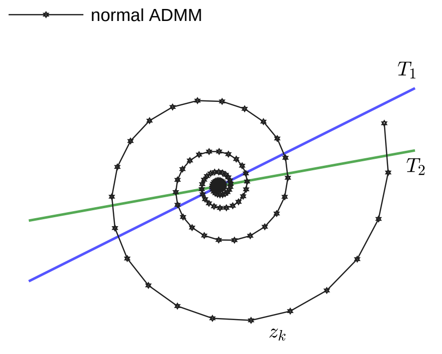
Intuition: inertial-ADMM accelerates if z_k is moving along a straight path...

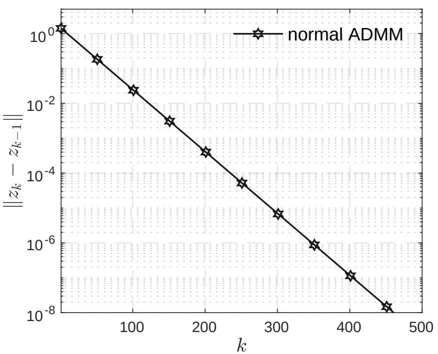
Failure of inertial-ADMM

Find $x \in T_1 \cap T_2$. 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:**



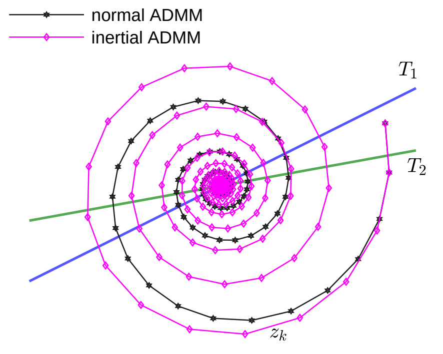


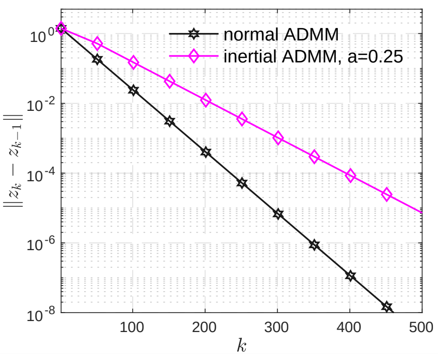
Failure of inertial-ADMM

Find $x \in T_1 \cap T_2$. 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 $a = 0.25$:**

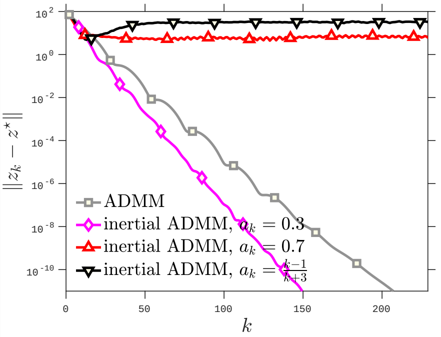




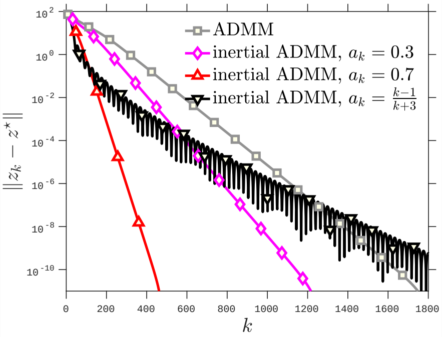
LASSO example:

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

$\gamma = \|K\|^2/10$



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



Eventual trajectory:

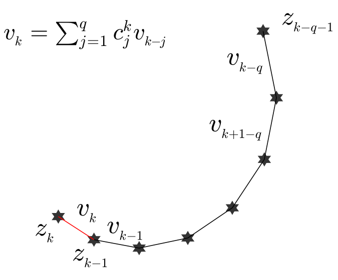
- Straight line when $\gamma > \|K\|^2$
- M_{ADMM} may have complex leading eigenvalue if $\gamma \leqslant \|K\|^2$.

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

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

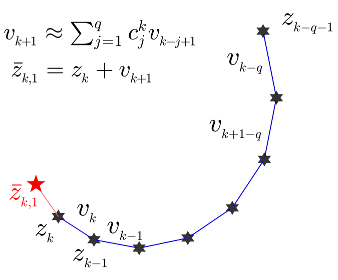
- Fit the past directions $v_{k-1}, \ldots, v_{k-q}$ to the latest direction v_k:

$$c^k \stackrel{\text{def.}}{=} \operatorname{argmin}_{c \in \mathbb{R}^q} \big\| \textstyle\sum_{j=1}^{q} c_j v_{k-j} - v_k \big\|^2.$$



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:
$$c^k \overset{\text{def.}}{=} \text{argmin}_{c \in \mathbb{R}^q} \| \textstyle\sum_{j=1}^{q} c_j v_{k-j} - v_k \|^2.$$
- Let $\bar{z}_{k,1} \overset{\text{def.}}{=} z_k + \sum_{j=1}^{q} c_j^k v_{k-j+1}$.



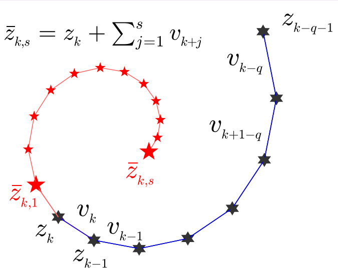
Idea: Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, define $\{v_{k-j} \stackrel{\text{def.}}{=} z_{k-j} - z_{k-j-1}\}_{j=0}^{q}$.

- Fit the past directions $v_{k-1}, \dots, v_{k-q}$ to the latest direction v_k:

$$c^k \stackrel{\text{def.}}{=} \operatorname{argmin}_{c \in \mathbb{R}^q} \left\| \sum_{j=1}^{q} c_j v_{k-j} - v_k \right\|^2.$$

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

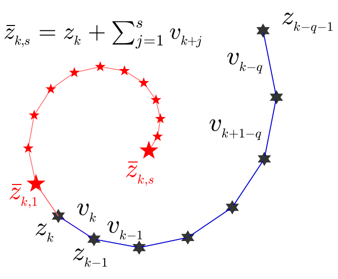
Repeat on $\{z_{k-j}\}_{j=0}^{q} \cup \{\bar{z}_{k,1}\}$ and so on.



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

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

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

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

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

Repeat: For $k \geqslant 1$

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

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

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

If $\mathrm{mod}(k, q+2) \neq 0$: $\bar{z}_k = z_k$.

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 $\varepsilon_k = \mathcal{O}(|\lambda_{q+1}|^k)$ where λ_{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

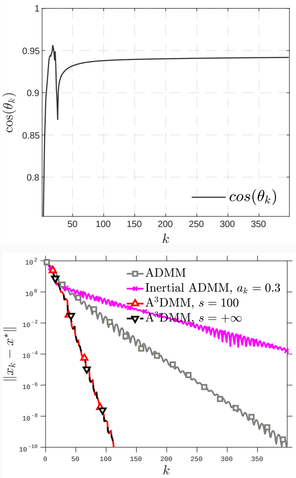
Implementation:

- Typically set $q \leqslant 10$.
- 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.

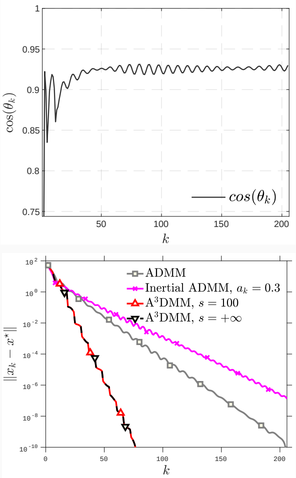
Experiment: 2 non-smooth terms

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

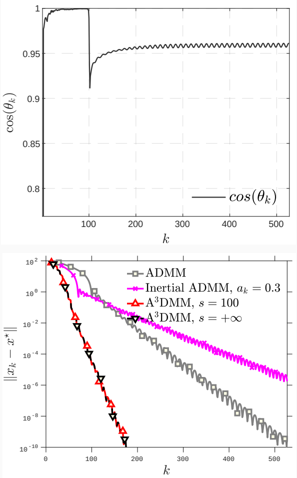
$$\min_{x,y \in \mathbb{R}^n} \ R(x) + \iota_\Omega(y) \quad \text{such that} \quad x - y = 0.$$



ℓ_1-norm

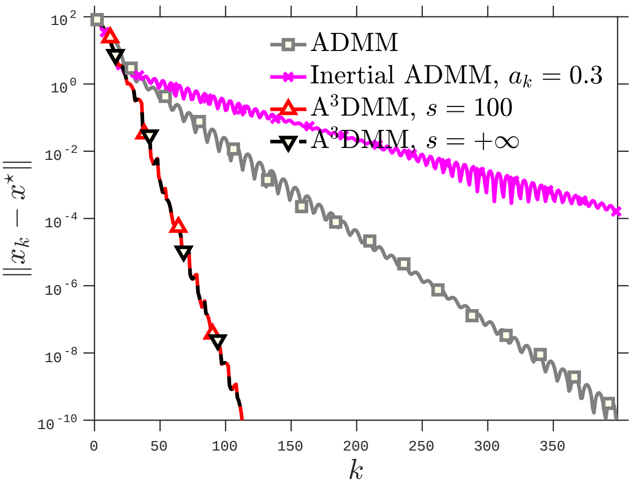


$\ell_{1,2}$-norm



Nuclear norm

Experiment: 2 non-smooth terms

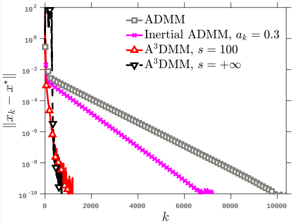


ℓ_1-norm

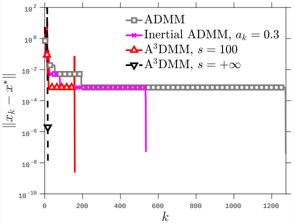
Inertial ADMM is **slower** than ADMM as eventual trajectory is a spiral.

The LASSO problem

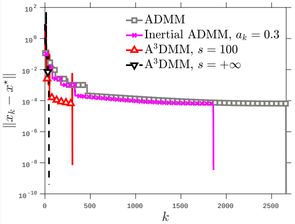
$$\min_{x,y\in\mathbb{R}^n} R(x) + \frac{1}{2}\|Ky - f\|^2 \quad \text{such that} \quad x - y = 0.$$



covtype

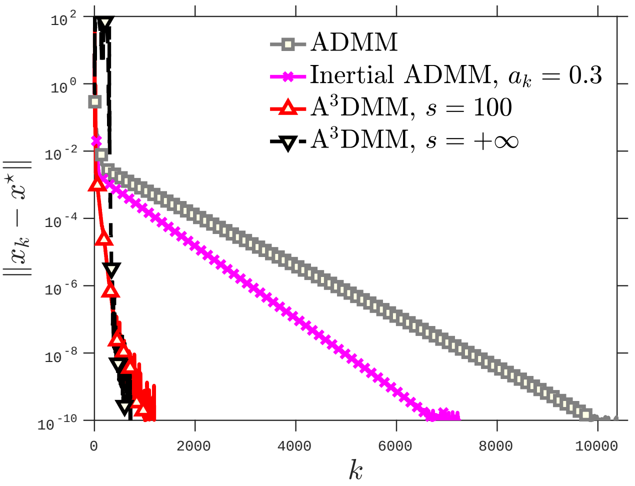


ijcnn1



phishing

Experiment: LASSO



covtype

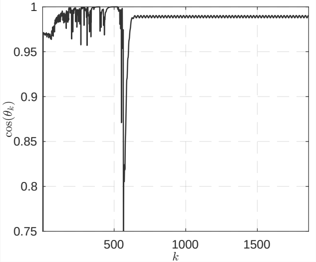
Inertial ADMM does accelerate, but A^{3}DMM is significantly faster.

Experiment: Total variation based image inpainting

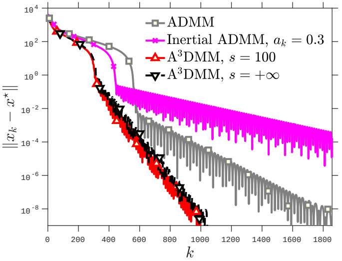
Let $\Omega \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^{n\times n} \;:\; P_{\mathcal{D}}(x) = f\}$, $P_{\mathcal{D}}$ randomly sets 50% pixels to zero and consider

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

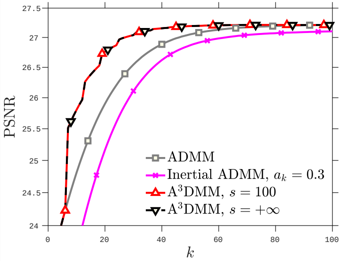
Angle $\{\theta_k\}$



$\|x_k - x^\star\|$



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 = 26.5448

Inertial ADMM, PSNR = 26.1096

Corrupted image

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

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