Trajectory of Alternating Direction Method of Multipliers and Adaptive Acceleration

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

A composite and constrained optimisation problem

Consider the following 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.

Augmented Lagrangian:

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

where $\gamma > 0$ and $\psi \in \mathbb{R}^p$ is the Lagrangian multiplier.

ADMM

The ADMM iterations are:

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

It is well known that ADMM is equivalent to applying the Douglas-Rachford (DR) iterations on the dual of $(\mathcal{P})$ and the equivalent DR iterates are

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

Moreover, there is a fixed-point operator $\mathcal{F}$ such that $z_k = \mathcal{F}(z_{k-1})$.

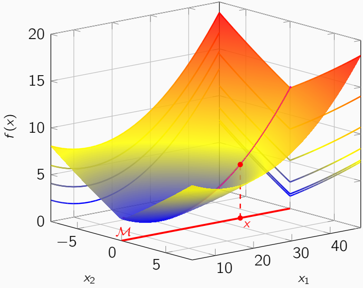
Partial smoothness

Definition [Lewis '05]: Let $R \in \Gamma_0(\mathbb{R}^n)$, R is partly smooth at x relative to a set $\mathcal{M}$ containing 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 $T_x \stackrel{\text{def.}}{=} \mathrm{par}\,(\partial R(x))^{\perp}$

Continuity: $\partial R : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is continuous along $\mathcal{M}$ near x



Examples:

- $\ell_1, \ell_{1,2}, \ell_\infty$-norm
- Nuclear norm
- Total variation

$\mathrm{par}(C)$: sub-space parallel to C, where $C \subset \mathbb{R}^n$ is a non-empty convex set.

Partial smoothness

It is known that under nondegeneracy conditions around the fixed points, if R and J are both partly smooth functions, then the behaviour of z_k is eventually **regular**.

Local linearisation [Liang, Fadili & Peyré '16]

There exists $K \in \mathbb{N}$ and a matrix M_{ADMM} such that for all $k \geqslant K$,

$$v_k = M_{\text{ADMM}} v_{k-1} + \varphi_{k-1}, \quad \text{where} \quad \varphi_{k-1} = o(\|v_{k-1}\|).$$

We will discuss the implications of this for the case where

- R and J are both non-smooth.
- At least one of R or J is smooth.

Partial smoothness and sequence trajectory

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

Two non-smooth terms

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

- $\varphi_k = 0$, M_{ADMM} is normal
- **Spiral trajectory:** $\cos(\theta_k) = \cos(\alpha) + \mathcal{O}(\eta^{2k})$ for some $\eta < 1$.

At least one smooth term

Suppose A is full rank square matrix and R is locally $\mathcal{C}^2$ around x^*. Then

- Eigenvalues of M_{ADMM} are all real-valued for $\gamma > \|(A^\top A)^{-\frac{1}{2}} \nabla^2 R(x^*)(A^\top A)^{-\frac{1}{2}}\|$.
- **Straight line trajectory:** $\cos(\theta_k) \to 1$.

Failure of inertial

Inertial ADMM

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

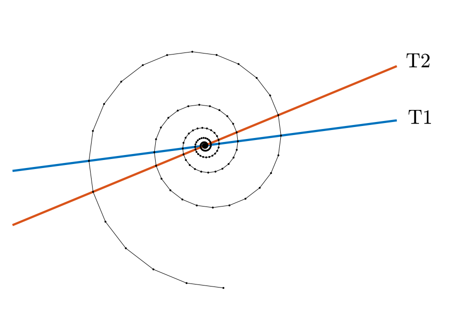
- Convergence is guaranteed for appropriate choice of a_k [Alvarez & Attouch '01].
- Acceleration guarantees are only available under additional assumptions such as Lipschitz smoothness and strong convexity [Pejcic & Jones '16, Kadkhodaie et al '15, França et al '18].

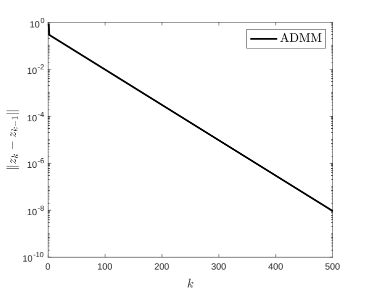
Failure of inertial

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:**



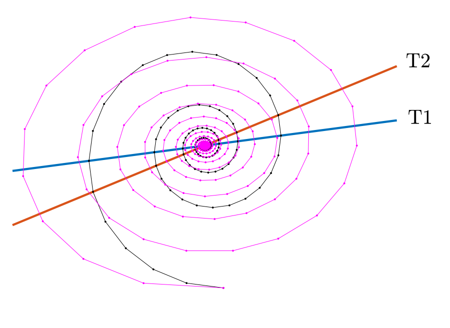


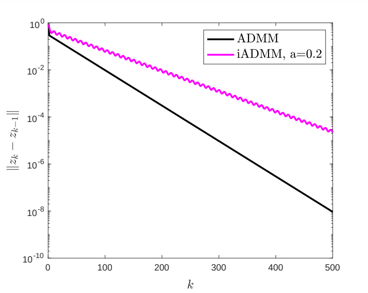
Failure of inertial

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 $a = 0.2$:**



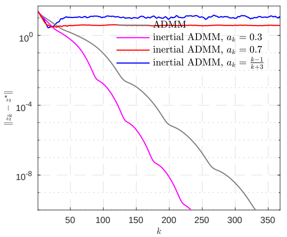


Failure of inertial

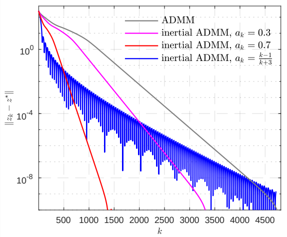
Consider the Lasso for a random Gaussian matrix $K \in \mathbb{R}^{m \times n}$ with $m < n$:

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

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



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



Eventual trajectory:

- Straight line when $\gamma > \|K\|^2$
- Linearisation matrix may have complex leading eigenvalue if $\gamma \leqslant \|K\|^2$.

Linear prediction

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

Idea

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

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

$$c_k \overset{\text{def.}}{=} \operatorname{argmin}_{c \in \mathbb{R}^q} \|V_{k-1}c - v_k\|^2, \quad \text{where} \quad V_{k-1} = [v_{k-1}, \ldots, v_{k-q}] \in \mathbb{R}^{n \times q}.$$

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

Linear prediction

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

Idea

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

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

$$c_k \stackrel{\text{def.}}{=} \operatorname{argmin}_{c \in \mathbb{R}^q} \|V_{k-1}c - v_k\|^2, \quad \text{where} \quad V_{k-1} = [v_{k-1}, \ldots, v_{k-q}] \in \mathbb{R}^{n \times q}.$$

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}.

Linear prediction

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

Idea

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

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

$$c_k \overset{\text{def.}}{=} \operatorname{argmin}_{c \in \mathbb{R}^q} \|V_{k-1}c - v_k\|^2, \quad \text{where} \quad V_{k-1} = [v_{k-1}, \ldots, v_{k-q}] \in \mathbb{R}^{n \times q}.$$

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}.

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

NB: $\bar{V}_{k,1} \overset{\text{def.}}{=} [(\bar{z}_{k,1} - z_k)|v_k| \cdots |v_{k-q+1}]$. The **$s$-step extrapolation** is

$$\bar{z}_{k,s} = z_k + \sum_{j=1}^{s} \left(\bar{V}_{k,j}\right)_{(:,1)} = z_{k+1} + \underbrace{V_k \left(\sum_{j=1}^{s} H(c_k)^j \right)_{(:,1)}}_{\mathcal{E}_{s,q}\left(\{z_{k-j}\}_{j=0}^{q}\right)}$$

Linear prediction

Adaptive Acceleration for ADMM

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

Repeat: For $k \geqslant 1$

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

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

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

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

Remarks

- Typically set $q \leqslant 10$.
- When $\rho(C_k)) < 1$, $\sum_{i=1}^{s} C_k^i = \begin{cases} (C_k - C_k^{s+1})(\mathrm{Id} - C_k)^{-1} & s < \infty \\ (\mathrm{Id} - C_k)^{-1} - \mathrm{Id} & s = +\infty \end{cases}$.
- Extra memory cost of $n \times (q+1)$ (storing V_k).
- Extra computation cost of $q^2 n$ every $(q+2)$ iterations.
- One could also extrapolate $\{x_k, y_k\}$ simultaneously. But this would require extra storage of past directions.

Theoretical guarantees

Global convergence:

- If $z_k = \mathcal{F}(z_{k-1})$ converges to fixed point z_*, then iterates $z_k = \mathcal{F}(z_{k-1} + \varepsilon_{k-1})$ also converge to z_*.
- Convergence is therefore guaranteed by appropriate choice of a_k.

Theoretical guarantees

Global convergence:

- If $z_k = \mathcal{F}(z_{k-1})$ converges to fixed point z_*, then iterates $z_k = \mathcal{F}(z_{k-1} + \varepsilon_{k-1})$ also converge to z_*.
- Convergence is therefore guaranteed by appropriate choice of a_k.

Local acceleration: Let $v_k \stackrel{\text{def.}}{=} z_k - z_{k-1}$ and assume that $v_k = Mv_{k-1}$.

- Coefficients fitting error: $\varepsilon_k \stackrel{\text{def.}}{=} \min_c \|V_{k-1}c - v_k\|$.
- For $s \in \mathbb{N}$, $\|\bar{z}_{k,s} - z^*\| \leqslant \|z_{k+s} - z^*\| + B_s\varepsilon_k$. If $\rho(M) < 1$ and $\rho(C_k) < 1$, then B_s is uniformly bounded in s.

Coefficients fitting error

Suppose that M is diagonalisable. Denote its distinct eigenvalues by $\left(\lambda_j\right)_j$ and order them in decreasing order.

- Asymptotic bound (fixed q and let $k \to +\infty$): $\varepsilon_k = \mathcal{O}(|\lambda_{q+1}|^k)$.
- Non-asymptotic bound (fixed q and k): Suppose that $\lambda(M)$ is real-valued and contained in the interval $[\alpha, \beta]$ with $-1 < \alpha < \beta < 1$, then $\varepsilon_k \lesssim \beta^{k-q} \left(\frac{\sqrt{\eta}-1}{\sqrt{\eta}+1}\right)^q$, where $\eta = \frac{1-\alpha}{1-\beta}$.

Remark: There is perfect linearisation for all k sufficiently large in the case where R and J are both polyhedral. Local acceleration is guaranteed with the choice of $q = 2$.

Previous works

The topic of convergence acceleration is a well-established field in numerical analysis.

1927 Aitkin's Δ-process.

1965 Andersen's acceleration.

1970's Vector extrapolation techniques such as minimal polynomial extrapolation (MPE) and reduced rank extrapolation (RRE) [Sidi '17].

2016 - Regularized non-linear acceleration (RNA) is a regularised version of RRE introduced by [Scieur et al '16].

Previous works

The topic of convergence acceleration is a well-established field in numerical analysis.

1927 Aitkin's Δ-process.

1965 Andersen's acceleration.

1970's Vector extrapolation techniques such as minimal polynomial extrapolation (MPE) and reduced rank extrapolation (RRE) [Sidi '17].

2016 - Regularized non-linear acceleration (RNA) is a regularised version of RRE introduced by [Scieur et al '16].

Relations to our work:

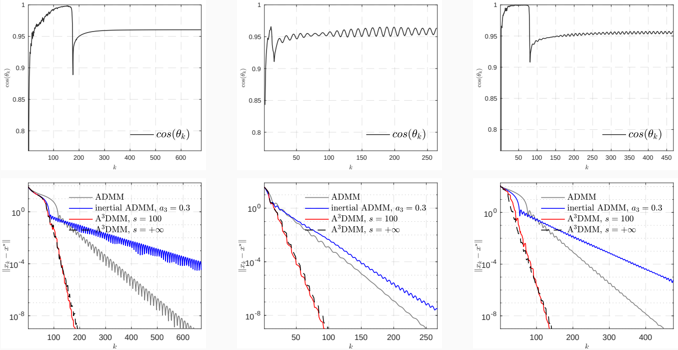
1. Based on the notion of minimal polynomials, MPE aims to compute the limit point of $\{z_j\}_{j\in\mathbb{N}}$ given points $\{z_j\}_{j=0}^{q+1}$ by computing $\bar{z} = \sum_{j=0}^{q} c_j z_j$.
2. Linear prediction with infinite step is the same as MPE shifted by one point $\bar{z}_\infty = \sum_{j=0}^{q} c_j z_{j+1}$.

Our formulation gives an alternative viewpoint on MPE, specific to nonsmooth optimisation.

Experiment: Affine constrained minimisation

Consider the basis pursuit 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 $\quad$ $\ell_{1,2}$ $\quad$ Nuclear

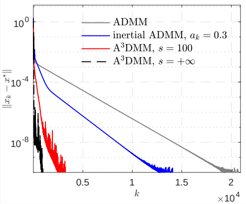
- Both functions are nonsmooth (and both are polyhedral for $R = \ell_1$).
- Inertial ADMM is slower than ADMM as eventual trajectory is a spiral.

Experiment: Lasso

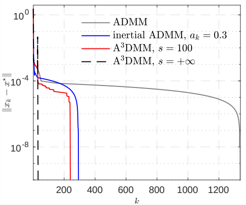
Consider the Lasso problem

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

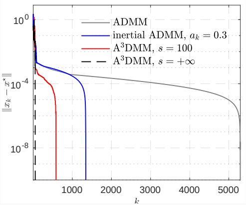
covtype



ijcnn1



phishing



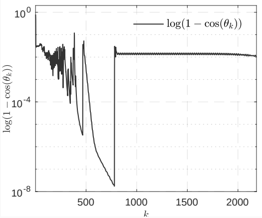
Although inertial ADMM provides acceleration, A^{3}DMM is significantly faster.

Experiment: Total variation based image inpainting

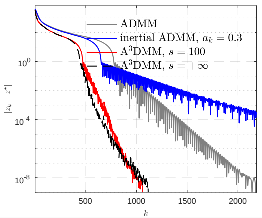
Let $\Omega \stackrel{\text{def.}}{=} \{x \in \mathbb{R}^{n\times n} \;;\; P_{\mathcal{D}}(x) = f\}$, $P_{\mathcal{D}}$ randomly sets 50% pixels to zero and consider

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

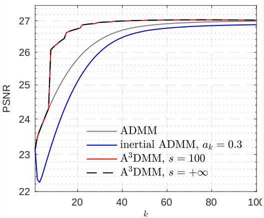
Angle $\{\theta_l\}$



$\|z_k - z^*\|$



PSNR



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

Summary of contributions

Trajectory analysis

Under the assumption that R and J are partly smooth functions, $\{z_k\}_k$ eventually settles onto a regular trajectory. In particular:

1. When both R and J are locally polyhedral (hence non-smooth) around the fixed point, z_k eventually moves along a spiral.
2. When at least one of R or J is smooth, the trajectory of z_k depends on γ and can be either a spiral or a straight line.

An adaptive acceleration scheme 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.
- This provides an alternative geometric interpretation of vector extrapolation techniques such as MPE and RRE.