

Globally Convergent Type-I Anderson Acceleration for Non-Smooth Fixed-Point Iterations

Theo Koppenhöfer

Lund

April 9, 2023

Table of contents

An introductory example

AA-II

AA-II

AA-I

Modifications to AA-I

Powell-type regularisation

Restarting iteration

Safeguarding steps

Convergence result

Numerical experiments

Regularised logistic regression

Elastic net regression

Summary

Sources

The problem setting

Problem (find fixed point)

Find a fixed point $x \in \mathbb{R}^n$ of $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$, i.e. $x = f(x)$.

or equivalently

Problem (find a zero)

Find a zero $x \in \mathbb{R}^n$ of $g = \text{Id} - f$, i.e. $0 = g(x)$.

We also assume

- ▶ f is nonexpansive, i.e. $\|f(x) - f(y)\| \leq \|x - y\|$
- ▶ n is large \rightarrow matrix-free
- ▶ ∇f is unknown \rightarrow no Newton
- ▶ cost of evaluation of f is high \rightarrow no line search
- ▶ noisy problem \rightarrow no finite difference derivatives

Algorithm 1: General AA

Input: $x^0 \in \mathbb{R}^n, f: \mathbb{R}^n \rightarrow \mathbb{R}^n$

for $k = 0, 1, \dots$ **do**

 Choose $m_k \in \{0, \dots, k\}$;

 Choose $\alpha^k \in \mathbb{R}^{m_k}$ such that $\sum_i \alpha_i^k = 1$;

$f_k = f(x_k)$ $x_{k+1} = \sum_i \alpha_i^k f_{k-m_k+i}$;

end

Algorithm 2: General AA

Input : $x^0 \in \mathbb{R}^n, f: \mathbb{R}^n \rightarrow \mathbb{R}^n$

for $k = 0, 1, \dots$ **do**

 Choose $m_k \in \{0, \dots, k\}$;

 Choose $\alpha^k \in \mathbb{R}^{m_k}$ such that $\sum_i \alpha_i^k = 1$ and such that α
 minimises $\|\sum_i \alpha_i^k g^i\|_2$;

$f_k = f(x_k)$;

$x_{k+1} = \sum_i \alpha_i^k f_{k-m_k+i}$;

end

Define residual $g = \text{Id} - f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $g_k = g(x_k)$. Choose $\alpha \in \mathbb{R}^{m_k}$ such that it minimises

$$\left\| \sum_i \alpha_i^k g_i \right\|_2$$

and

$$\sum_i \alpha_i^k = 1.$$

It can be shown that then

$$x_{k+1} = \sum_i \alpha_i^k f(x_{k-m_k+i}) = x_k - H_k g_k$$

for some $H_k \in \mathbb{R}^{n \times m_k}$ such that H_k minimises $\|H_k - \text{Id}\|_F$.

Algorithm 3: AA-I

Input : $x^0 \in \mathbb{R}^n$ and $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$

Set $H_0 = \text{Id}$, $x_1 = f(x_0)$.

for $k = 0, 1, \dots$ **do**

 Set $g_k = g(x_k)$, $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$.

 Set $\hat{s}_{k-1} = s_{k-1} - \sum_{i=0}^{k-2} \frac{\hat{s}_i^\top s_{k-1}}{\|\hat{s}_i\|^2} s_i$.

 Set $H_k = H_{k-1} + \frac{(s_{k-1} - H_{k-1} y_{k-1}) s_{k-1}^\top H_{k-1}}{\hat{s}_{k-1}^\top H_{k-1} y_{k-1}}$ and $x_{k+1} = x_k - H_k g_k$.

end

Powell-type regularisation

Note that B_k may be singular. To solve this set

$$\tilde{y}_k = \theta_k y_k + (1 - \theta_k) B_k s_k$$

where

with

$$\theta_k = \phi_{\bar{\theta}}(\eta_k)$$
$$\phi_{\bar{\theta}}(\eta) = \begin{cases} \frac{1 - \operatorname{sgn}(\eta)\bar{\theta}}{1 - \eta} & \text{if } |\eta| < \bar{\theta} \\ 1 & \text{else} \end{cases} \quad \text{and} \quad \eta_k = \frac{\hat{s}_k^\top H_k y_k}{\|\hat{s}_k\|^2}$$

One can obtain

Lemma (Powell-type regularisation)

Let $s_k \in \mathbb{R}^n$, $B_0 = \text{Id}$, and inductively

$$B_{k+1} = B_k + \frac{(\tilde{y}_k - B_k s_k) \hat{s}_k^\top}{\hat{s}_k^\top s_k}$$

with \hat{s}_k and \tilde{y}_k defined as before. If this is well-defined then $|\det(B_k)| \geq \theta^k > 0$ and B_k is invertible.

Proof.

See [1, Lemma 2].



Algorithm 4: AA-I with Powell-like-regularisation

Input : $x^0 \in \mathbb{R}^n$, $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $\bar{\theta} \in (0, 1)$

Set $H_0 = \text{Id}$, $x_1 = f(x_0)$.

for $k = 0, 1, \dots$ **do**

Set $g_k = g(x_k)$, $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$.

Set $\hat{s}_{k-1} = s_{k-1} - \sum_{i=0}^{k-2} \frac{\hat{s}_i^\top s_{k-1}}{\|\hat{s}_i\|^2} s_i$.

Set $\eta_{k-1} = \frac{\hat{s}_{k-1}^\top H_{k-1} y_{k-1}}{\|\hat{s}_{k-1}\|^2}$, $\theta_{k-1} = \phi_{\bar{\theta}}(\eta_{k-1})$ and

$\tilde{y}_{k-1} = \theta_{k-1} y_{k-1} - (1 - \theta_{k-1}) g_{k-1}$.

Set $H_k = H_{k-1} + \frac{(s_{k-1} - H_{k-1} \tilde{y}_{k-1})}{\hat{s}_{k-1}^\top H_{k-1} \tilde{y}_{k-1}}$ and $x_{k+1} = x_k - H_k g_k$.

end

Restarting iteration

Note that

$$B_{k+1} = B_k + \frac{(\tilde{y}_k - B_k s_k) \hat{s}_k^\top}{\hat{s}_k^\top s_k}$$

is ill-defined iff $\|\hat{s}_k\|^2 = \hat{s}_k^\top s_k = 0$, i.e. $\hat{s}_k = 0$. This can occur for $m_k > n$ as we then have $\hat{s}_k = 0$ by linear dependence. If we reset $m_k = 0$ if $m_k = m + 1$ or $\|\hat{s}_{k-1}\| < \tau \|s_{k-1}\|$ for some $\tau \in (0, 1)$ then

$$g_k \neq 0 \implies s_k = -B_k g_k \neq 0 \implies \hat{s}_k \neq 0.$$

Lemma (Restarting iteration)

If we additionally choose m_k by the rule above we have

$$\|B_k\| \leq 3 \left(\frac{1 + \bar{\theta} + \tau}{\tau} \right)^m - 2.$$

Proof.

See [1, Lemma 3].



Algorithm 5: AA-I with Powell-like-regularisation and Restarting

Input : $x^0 \in \mathbb{R}^n, f: \mathbb{R}^n \rightarrow \mathbb{R}^n, m \in \mathbb{N}$ and $\bar{\theta}, \tau \in (0, 1)$

Set $H_0 = \text{Id}, x_1 = f(x_0), m_0 = 0$.

for $k = 0, 1, \dots$ **do**

 Set $g_k = g(x_k), m_k = m_{k-1} + 1, s_{k-1} = x_k - x_{k-1}$ and

$y_{k-1} = g_k - g_{k-1}$.

 Set $\hat{s}_{k-1} = s_{k-1} - \sum_{i=k-m_k}^{k-2} \frac{\hat{s}_i^\top s_{k-1}}{\|\hat{s}_i\|^2} s_i$.

if $m_k = m + 1$ **or** $\|\hat{s}_{k-1}\| < \tau \|s_{k-1}\|$ **then**

 Set $m_k = 0, \hat{s}_{k-1} = s_{k-1}$ and $H_{k-1} = \text{Id}$.

end

 Set $\eta_{k-1} = \frac{\hat{s}_{k-1}^\top H_{k-1} y_{k-1}}{\|\hat{s}_{k-1}\|^2}, \theta_{k-1} = \phi_{\bar{\theta}}(\eta_{k-1})$ and

$\tilde{y}_{k-1} = \theta_{k-1} y_{k-1} - (1 - \theta_{k-1}) g_{k-1}$.

 Set $H_k = H_{k-1} + \frac{(s_{k-1} - H_{k-1} \tilde{y}_{k-1})}{\hat{s}_{k-1}^\top H_{k-1} \tilde{y}_{k-1}}$ and $x_{k+1} = x_k - H_k g_k$.

end

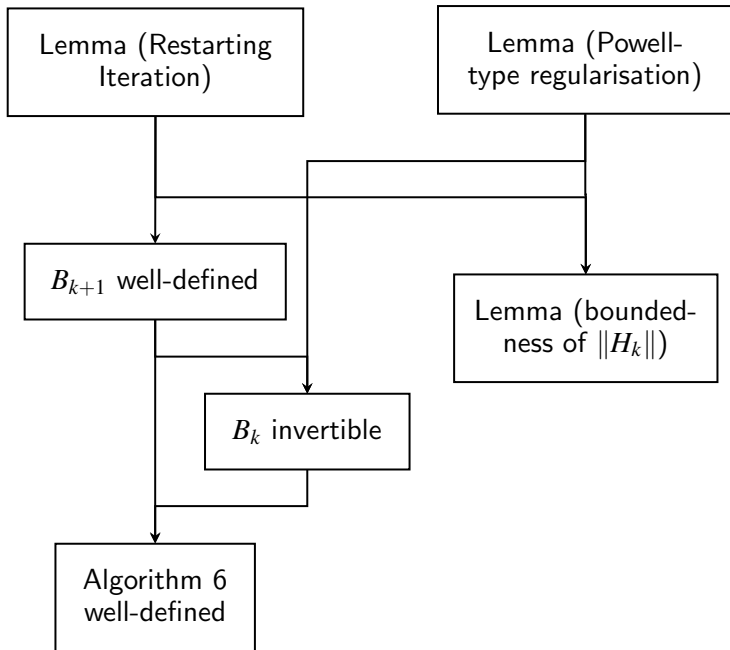
Lemma (bound on $\|H_k\|_2$)

In algorithm 5 we have that

$$\|H_k\|_2 \leq \frac{1}{\bar{\theta}^m} \left(3 \left(\frac{1 + \bar{\theta} + \tau}{\tau} \right)^m - 2 \right)^{n-1}.$$

Proof.

This follows from Lemma (Restarting iteration) and Lemma (Powell-type regularisation). □



Safeguarding steps

To guarantee the decrease in $\|g_k\|$ one can interleave the AA-I steps with Krasnosel'skii-Mann steps which are given by

$$x_{k+1} = (1 - \alpha)x_k + \alpha f(x_k)$$

for some fixed $\alpha \in (0, 1)$.

Algorithm 6: AA-I with Powell-like-regularisation, Restarting and Safeguarding

Input : $x^0 \in \mathbb{R}^n$, $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$, $m \in \mathbb{N}$, $\bar{\theta}, \tau, \alpha \in (0, 1)$, safe-guarding constants $D, \varepsilon > 0$

Set $H_0 = \text{Id}$, $x_1 = \tilde{x}_1 = f(x_0)$, $m_0 = n_{AA} = 0$ and $\bar{U} = \|g_0\|_2$.

for $k = 0, 1, \dots$ **do**

 Set $g_k = g(x_k)$, $m_k = m_{k-1} + 1$, $s_{k-1} = \tilde{x}_k - x_{k-1}$ and $y_{k-1} = g(\tilde{x}_k) - g_{k-1}$.

 Set $\hat{s}_{k-1} = s_{k-1} - \sum_{i=k-m_k}^{k-2} \frac{\hat{s}_i^\top s_{k-1}}{\|\hat{s}_i\|^2} s_i$.

if $m_k = m + 1$ **or** $\|\hat{s}_{k-1}\| < \tau \|s_{k-1}\|$ **then**

 Set $m_k = 0$, $\hat{s}_{k-1} = s_{k-1}$ and $H_{k-1} = \text{Id}$.

end

 Set $\eta_{k-1} = \frac{\hat{s}_{k-1}^\top H_{k-1} y_{k-1}}{\|\hat{s}_{k-1}\|^2}$, $\theta_{k-1} = \phi_{\bar{\theta}}(\eta_{k-1})$ and

$\tilde{y}_{k-1} = \theta_{k-1} y_{k-1} - (1 - \theta_{k-1}) g_{k-1}$.

 Set $H_k = H_{k-1} + \frac{(s_{k-1} - H_{k-1} \tilde{y}_{k-1})}{\hat{s}_{k-1}^\top H_{k-1} \tilde{y}_{k-1}}$ and $\tilde{x}_{k+1} = x_k - H_k g_k$.

if $\|g_k\| \leq D \bar{U} (n_{AA} + 1)^{-(1+\varepsilon)}$ **then**

 Set $x_{k+1} = \tilde{x}_{k+1}$ and $n_{AA} = n_{AA} + 1$.

else

 Set $x_{k+1} = (1 - \alpha)x_k + \alpha f(x_k)$

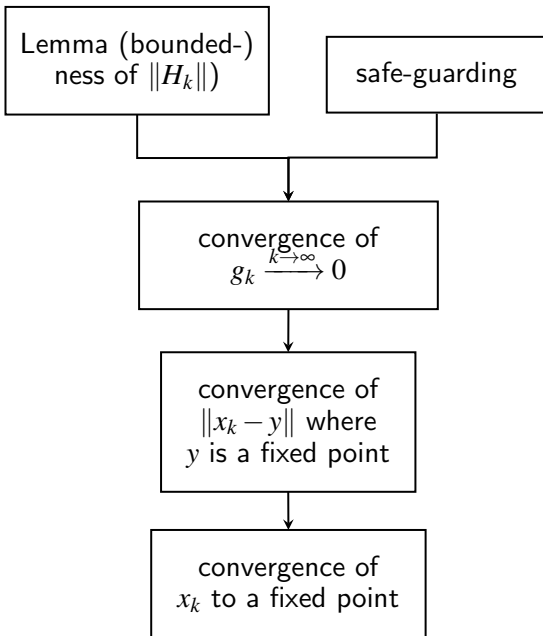
end

end

Convergence result

Theorem (Convergence)

Let x_k be generated by algorithm 6 then $x_k \xrightarrow{k \rightarrow \infty} x_$ and $f(x_*) = x_*$ is a fixed point.*



Regularised logistic regression

We take $x \in \mathbb{R}^{2000 \times 500}$, $y \in \mathbb{R}^{2000}$ from the UCI Madelon dataset [2].
The aim is to minimise

$$F(\theta) = \frac{1}{2000} \sum_i \log(1 + \sum_j y_i x_{ij} \theta_j) + \frac{\lambda}{2} \|\theta\|^2$$

with gradient descent, i.e.

$$f: \mathbb{R}^{500} \rightarrow \mathbb{R}^{500}, \quad \theta \mapsto \theta - \alpha \nabla F(\theta)$$

for some α .

residual norms for the problem GD

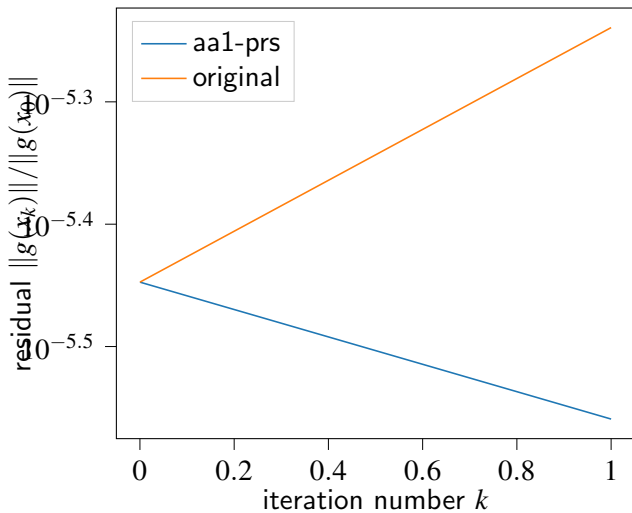


Figure: Residual norms for the logistic regression problem.

Facility location


The aim is to minimise

$$F: \mathbb{R}^{300} \rightarrow \mathbb{R}, \quad y \mapsto \sum_{i=1}^{500} \|y - c_i\|$$

for $c_i \in \mathbb{R}^{300}$ with sparsity 0.01. This can lead to the formulation

$$\tilde{f}: \mathbb{R}^{500 \times 300} \rightarrow \mathbb{R}^{500 \times 300}, \quad z \mapsto \left(z_i + 2 \langle x \rangle - x_i - \langle z \rangle \right)_i$$

with


$$\langle x \rangle = \frac{1}{500} \sum_i x_i \quad x_i = \text{prox}_{\|\cdot\|} (z_i + c_i) - c_i$$

and

$$\text{prox}_{\|\cdot\|} (v) = \left(1 - \frac{1}{\|v\|} \right)_+ v.$$

residual norms for the problem CO

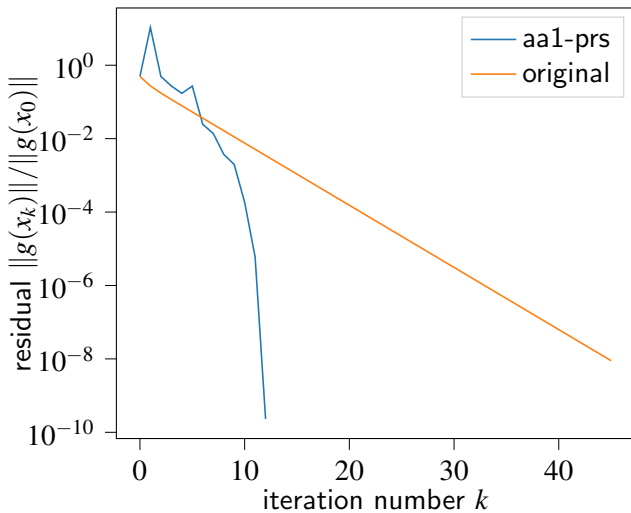


Figure: Residual norms for the facility location problem.

Elastic net regression

Our aim is to minimise

$$F: \mathbb{R}^{1000} \rightarrow \mathbb{R}, \quad x \mapsto \frac{1}{2} \|Ax - b\|^2 + \mu \left(\frac{1}{4} \|x\|^2 + \frac{1}{2} \|x\|_1 \right)$$

with $A \in \mathbb{R}^{500 \times 1000}$, $b \in \mathbb{R}^{500}$ and some $\mu \in \mathbb{R}$. From the Iterative Shrinkage-Thresholding Algorithm one obtains

$$f: \mathbb{R}^{1000} \rightarrow \mathbb{R}^{1000}, \quad x \mapsto S_{\alpha\mu/2} \left(x - \alpha \left(A^\top (Ax - b) + \frac{\mu}{2} x \right) \right)$$

with shrinkage operator

$$S_\kappa(x) = (\operatorname{sgn}(x_i)(|x_i| - \kappa)_+)_i$$

and some $\alpha \in \mathbb{R}$.

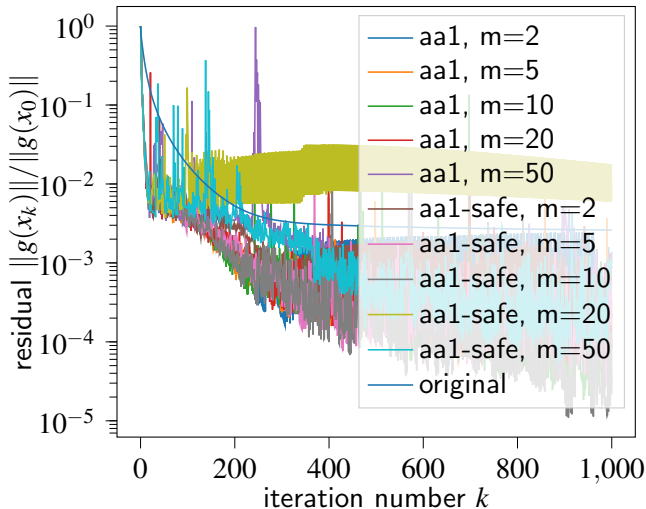


Figure: Residual norms for the elastic net regression problem.

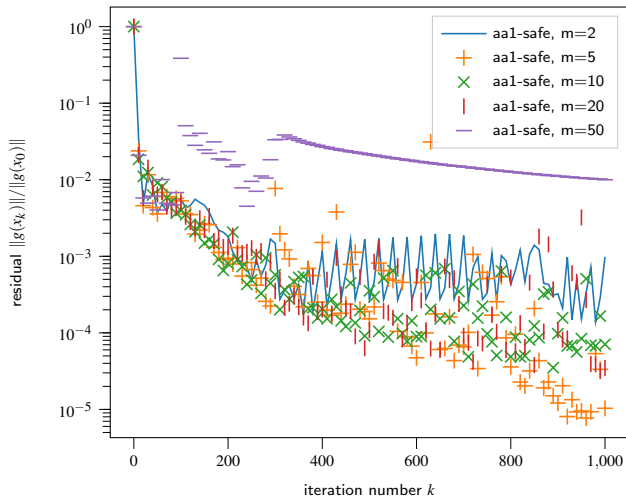


Figure: Residual norms for the elastic net regression problem.

Summary

Sources I

- [1] J. Zhang, B. O'Donoghue, and S. Boyd, "Globally convergent type-I Anderson acceleration for nonsmooth fixed-point iterations," *SIAM J. Optim.*, vol. 30, no. 4, pp. 3170–3197, 2020, ISSN: 1052-6234. DOI: 10.1137/18M1232772. [Online]. Available: <https://doi-org.ludwig.lub.lu.se/10.1137/18M1232772>.
- [2] I. Guyon. (2004), Madelon data set, [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Madelon>.
- [3] H.-r. Fang and Y. Saad, "Two classes of multiseant methods for nonlinear acceleration," *Numer. Linear Algebra Appl.*, vol. 16, no. 3, pp. 197–221, 2009, ISSN: 1070-5325. DOI: 10.1002/nla.617. [Online]. Available: <https://doi-org.ludwig.lub.lu.se/10.1002/nla.617>.

Thank you for your attention.

