Convergence Acceleration for Nonlinear Fixed-Point Methods

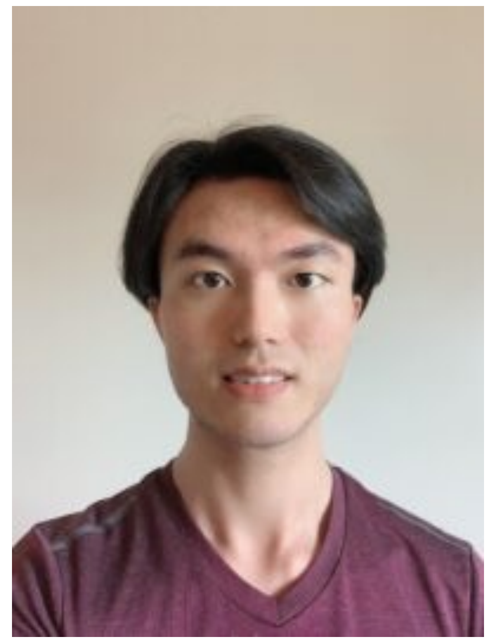
Hans De Sterck

University of Waterloo, Canada

UNIVERSITY OF WATERLOO

joint work with:

Yunhui He (postdoc) and Dawei Wang (graduate student)



(1) introduction

- we will consider fixed-point (FP) iterative methods to numerically compute approximate solutions of scientific computing or optimization problems

$$\boxed{\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)} \qquad \{\mathbf{x}_0, \mathbf{x}_1, \ldots\} \qquad \mathbf{x}^*$$

- for difficult (ill-conditioned) problems, (asymptotic) FP convergence may be (very) slow

- we will consider nonlinear acceleration methods to improve the (asymptotic) convergence, e.g., Anderson Acceleration (AA):

$$\boxed{\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i^{(k)} \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)}$$

hdesterck@uwaterloo.ca

3 applications

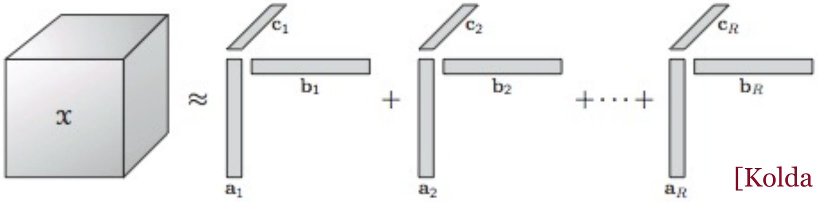
(A) solving linear equation systems: $A\mathbf{x} = \mathbf{b}$ $\quad A \in \mathbb{R}^{n\times n}, \quad \mathbf{x}, \mathbf{b} \in \mathbb{R}^n$

affine iteration: $\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$

(B) machine learning optimization problems

nonlinear iteration: $\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$ (Alternating Direction Method of Multipliers, ADMM)

(C) canonical tensor decomposition:



[Kolda and Bader, 2009]

nonlinear iteration: $\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$ (Alternating Least Squares, ALS)

(A) solving linear equations

solving linear equation systems:

- linear system $A\mathbf{x} = \mathbf{b} \qquad A \in \mathbb{R}^{n\times n}, \quad \mathbf{x}, \mathbf{b} \in \mathbb{R}^n$
- FP iteration: $\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$

  we choose: $\mathbf{q}(\mathbf{x}) = M\mathbf{x} + P\mathbf{b} \qquad M = I - PA$

  fixed point:
$$\begin{aligned} \mathbf{x} = \mathbf{q}(\mathbf{x}) &= (I - PA)\mathbf{x} + P\mathbf{b} \\ &\iff PA\mathbf{x} = P\mathbf{b} \\ &\iff A\mathbf{x} = \mathbf{b} \end{aligned}$$

- P is called the preconditiong matrix (Jacobi, Gauss-Seidel, ...)

solving linear equations

$$\mathbf{q}(\mathbf{x}) = M\mathbf{x} + P\mathbf{b}$$

$$\mathbf{q}(\mathbf{x}) = (I - PA)\mathbf{x} + P\mathbf{b}$$

$$PA\mathbf{x} = P\mathbf{b}$$

- FP method: $\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$
- error: $\mathbf{e}_k = \mathbf{x}^* - \mathbf{x}_k$
- error propagation equation:

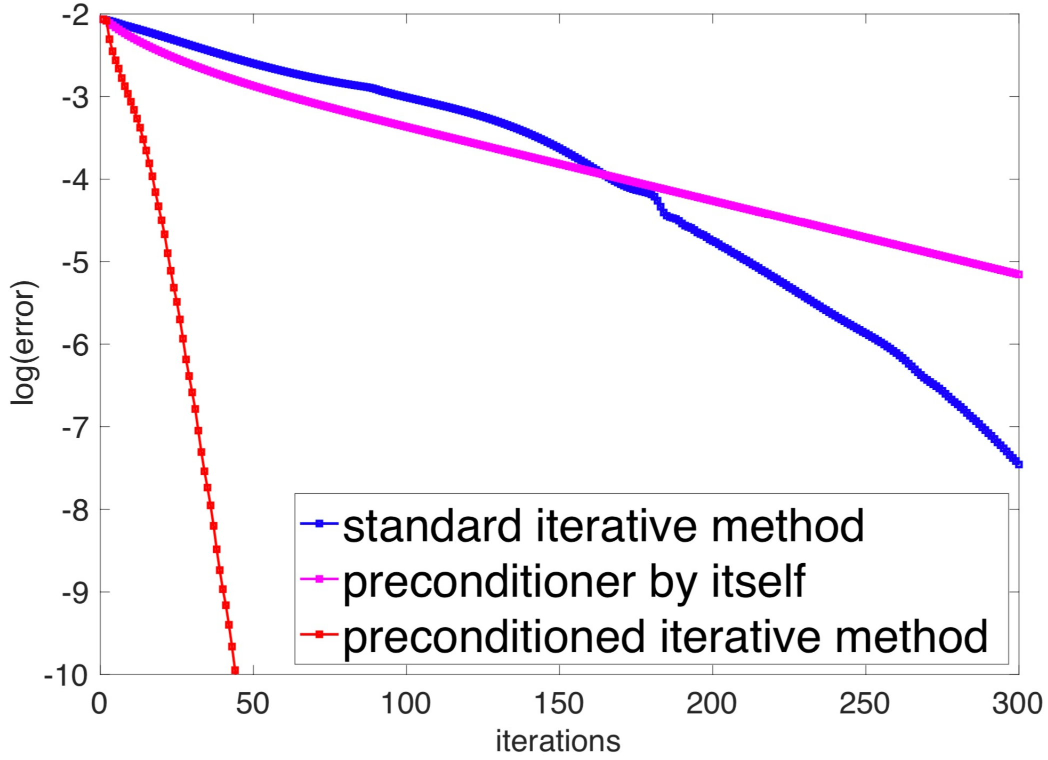
$$\mathbf{x}^* - \mathbf{x}_{k+1} = \mathbf{x}^* - (I - PA)\mathbf{x}_k - P\mathbf{b}$$

$$\boxed{\mathbf{e}_{k+1} = M\mathbf{e}_k}$$

- asymptotic convergence factor: $\rho(M)$
- in the nonlinear case: $\rho(\mathbf{q}'(\mathbf{x}^*))$

solving linear equations

- FP method may converge slowly: $\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$



$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$

(SSOR)

solving linear equations

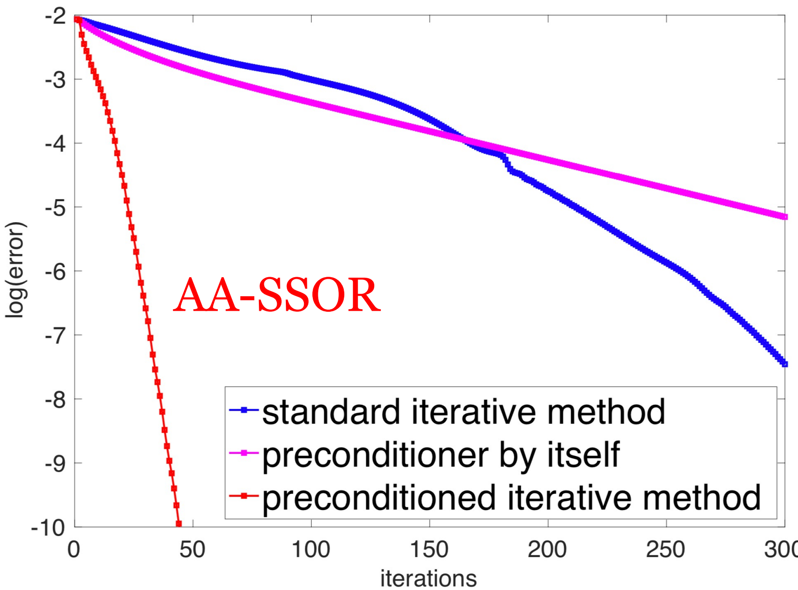
$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$$

- accelerate convergence by Anderson Acceleration (AA):

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i^{(k)} \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$

$$\mathbf{r}(\mathbf{x}_k) = \mathbf{x}_k - \mathbf{q}(\mathbf{x}_k)$$

$$\{\beta_i^{(k)}\} = \underset{\{\beta_i\}}{\operatorname{argmin}} \left\| \mathbf{r}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i \left(\mathbf{r}(\mathbf{x}_{k-i}) - \mathbf{r}(\mathbf{x}_{k-i-1})\right) \right\|^2$$



$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$$

(SSOR)

AA, P=I

UNIVERSITY OF WATERLOO

solving linear equations

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$$

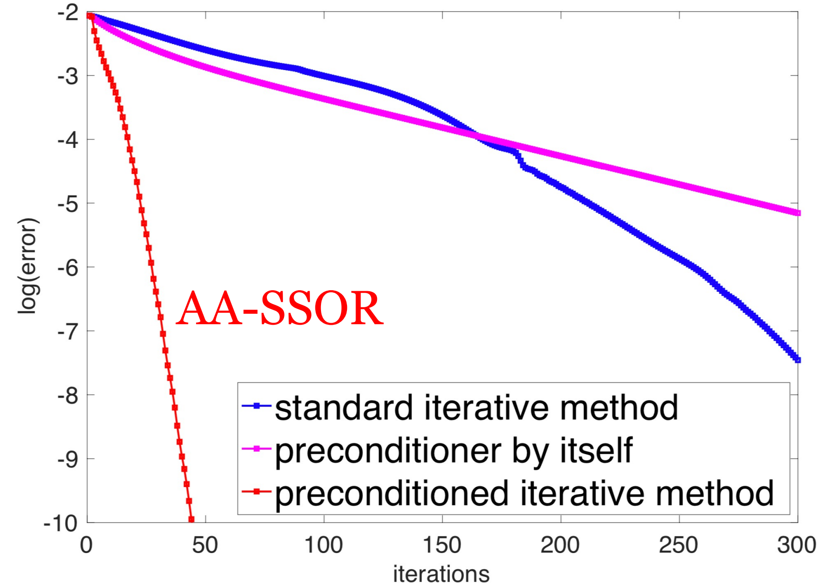
- accelerate convergence by Anderson Acceleration (AA):

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i^{(k)} \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$

$$\mathbf{r}(\mathbf{x}_k) = \mathbf{x}_k - \mathbf{q}(\mathbf{x}_k)$$

$$\{\beta_i^{(k)}\} = \underset{\{\beta_i\}}{\operatorname{argmin}} ||\mathbf{r}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i \left(\mathbf{r}(\mathbf{x}_{k-i}) - \mathbf{r}(\mathbf{x}_{k-i-1})\right)||^2$$

- in the linear case, AA is equivalent to the "Generalized minimal residual method" (GMRES)
- GMRES minimizes polynomials over a "Krylov space", which facilitates convergence analysis



solving linear equations

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$$

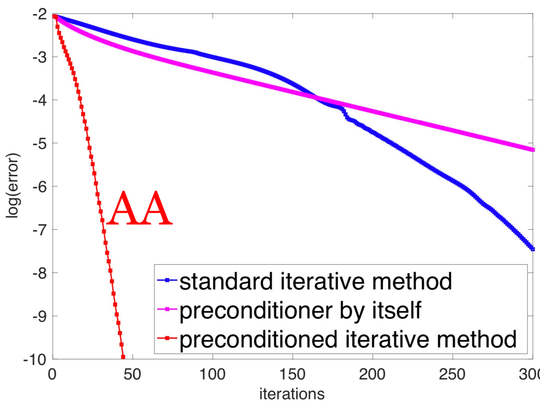
- GMRES convergence result:

Theorem 5.5

Let $A \in \mathbb{R}^{n\times n}$, nonsingular, be diagonalisable, $A = V\Lambda V^{-1}$. Then the residuals generated in the GMRES method satisfy

$$\frac{\|\vec{r}_i\|}{\|\vec{r}_0\|} \leq \kappa_2(V) \min_{p_i(x)\in\mathcal{P}_i} \max_{\Sigma(A)} |p_i(\lambda)|.$$

Here, $p_i(x)$ is a polynomial of degree at most i in $\mathcal{P}_i$, the set of polynomials of degree at most i which satisfy $p_i(0) = 1$. $\Sigma(A)$ is the eigenvalue spectrum of A, i.e., the set of eigenvalues of A.



$$\frac{\|r_k\|}{\|r_0\|} \leq c\rho^k$$

UNIVERSITY OF WATERLOO

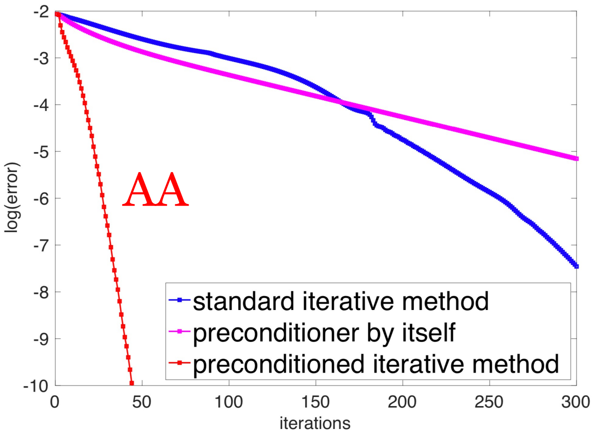
in more general cases, convergence analysis is hard …

- nonlinear FP iteration: $\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$
- Anderson Acceleration (AA) is usually done in a "windowed" fashion, window size m: $m_k = \min\{m, k\}$

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i^{(k)} \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$

$$\mathbf{r}(\mathbf{x}_k) = \mathbf{x}_k - \mathbf{q}(\mathbf{x}_k)$$

$$\{\beta_i^{(k)}\} = \underset{\{\beta_i\}}{\mathrm{argmin}} ||\mathbf{r}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i \left(\mathbf{r}(\mathbf{x}_{k-i}) - \mathbf{r}(\mathbf{x}_{k-i-1})\right)||^2$$



- we cannot rely on optimal polynomials, so convergence analysis is hard

(B) Alternating Direction Method of Multipliers – optimization for machine learning

- ADMM:

$$\min_{\mathbf{x},\mathbf{z}} f(\mathbf{x},\mathbf{z}) = f_1(\mathbf{x}) + f_2(\mathbf{z}),$$
$$\text{s.t.} \quad \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{b},$$

$$L_\rho(\mathbf{x},\mathbf{z},\mathbf{y}) = f_1(\mathbf{x}) + f_2(\mathbf{z}) + \mathbf{y}^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{b}) + \frac{\rho}{2}||\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{b}||_2^2$$

$$\begin{cases} \mathbf{x}_{k+1} = \text{argmin}_{\mathbf{x}} f_1(\mathbf{x}) + \frac{\rho}{2}||\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_k - \mathbf{b} + \mathbf{u}_k||_2^2, \\ \mathbf{z}_{k+1} = \text{argmin}_{\mathbf{z}} f_2(\mathbf{z}) + \frac{\rho}{2}||\mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z} - \mathbf{b} + \mathbf{u}_k||_2^2, \\ \mathbf{u}_{k+1} = \mathbf{u}_k + \mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z}_{k+1} - \mathbf{b}, \end{cases}$$

- ADMM as fixed-point method:

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$$

accelerating ADMM as a fixed-point method

- ADMM as fixed-point method:

$$\boxed{\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)}$$

$$\begin{cases} \mathbf{x}_{k+1} = \operatorname{argmin}_{\mathbf{x}} f_1(\mathbf{x}) + \frac{\rho}{2}||\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_k - \mathbf{b} + \mathbf{u}_k||_2^2, \\ \mathbf{z}_{k+1} = \operatorname{argmin}_{\mathbf{z}} f_2(\mathbf{z}) + \frac{\rho}{2}||\mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z} - \mathbf{b} + \mathbf{u}_k||_2^2, \\ \mathbf{u}_{k+1} = \mathbf{u}_k + \mathbf{A}\mathbf{x}_{k+1} + \mathbf{B}\mathbf{z}_{k+1} - \mathbf{b}, \end{cases}$$

- we consider problems where ADMM converges linearly, in particular, where $\mathbf{q}(\mathbf{x})$ is differentiable at $\mathbf{x}^*$

- we accelerate ADMM with Anderson Acceleration (AA):

$$\boxed{\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i^{(k)} \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)}$$

$$m_k = \min\{m, k\}$$

$$\{\beta_i^{(k)}\} = \underset{\{\beta_i\}}{\operatorname{argmin}} ||\mathbf{r}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i \left(\mathbf{r}(\mathbf{x}_{k-i}) - \mathbf{r}(\mathbf{x}_{k-i-1})\right)||^2$$

$$\mathbf{r}(\mathbf{x}_k) = \mathbf{x}_k - \mathbf{q}(\mathbf{x}_k)$$

ADMM accelerated by AA: LASSO example

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}||\mathbf{A}\mathbf{x}-\mathbf{b}||_2^2 + \lambda||\mathbf{z}||_1,$$
$$\text{s.t. } \mathbf{x}-\mathbf{z}=0.$$

(least absolute shrinkage and selection operator)

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$$

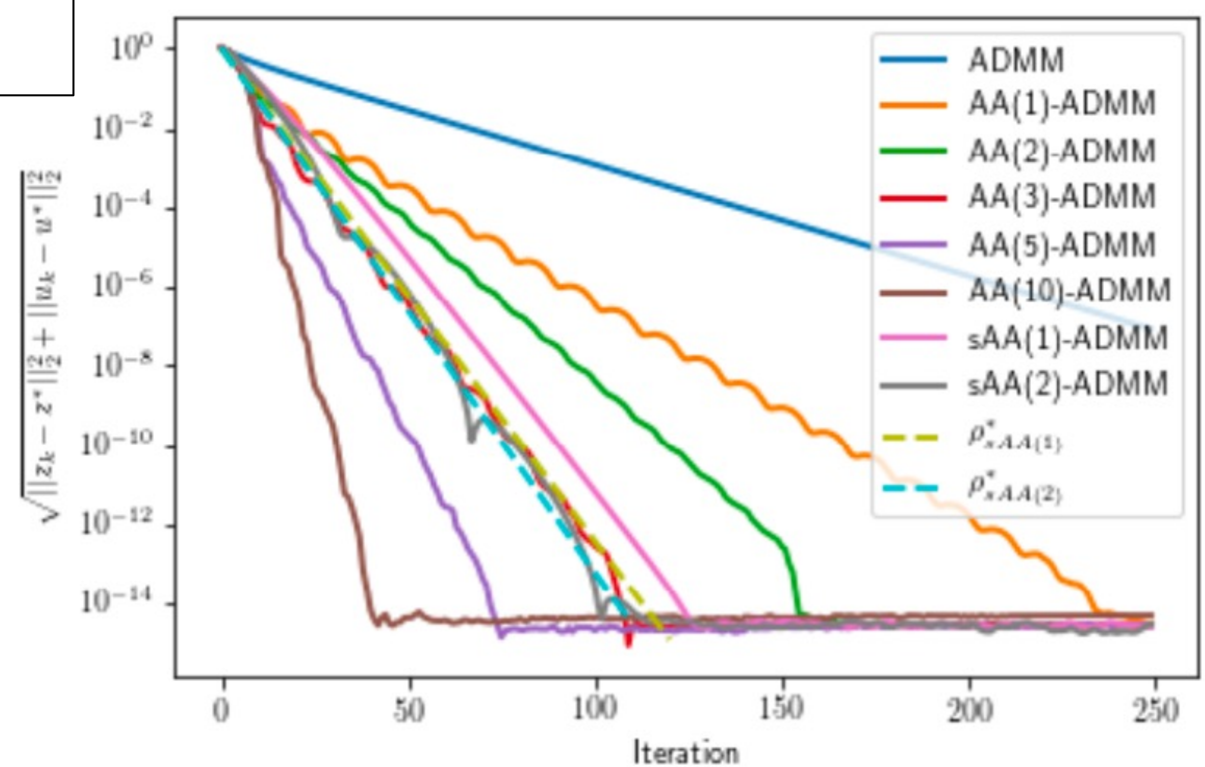
- convergence:

$$\rho_{ADMM,\mathbf{x}^*} = \rho(\mathbf{q}'(\mathbf{x}^*))$$

$$\rho_{AA-ADMM,\mathbf{x}^*} = \ ?$$

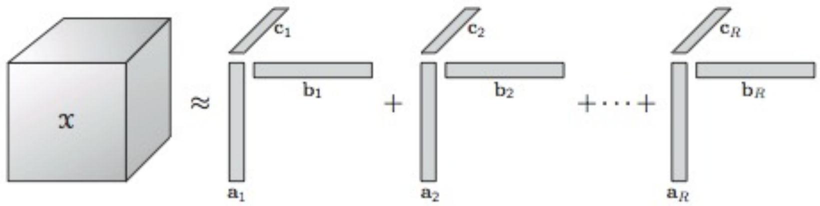
- our contribution:
  we consider optimal stationary version of AA (sAA)

$$\rho_{sAA-ADMM,\mathbf{x}^*} = \rho(\boldsymbol{\Psi}'(\mathbf{x}^*))$$



Zhang, J., Peng, Y., Ouyang, W., Deng, B.: Accelerating ADMM for efficient simulation and optimization. ACM Transactions on Graphics (TOG) **38**(6), 1–21 (2019)

(C) canonical tensor decomposition



OPTIMIZATION PROBLEM

given tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times \ldots \times I_N}$, find rank-$R$ canonical tensor $\mathcal{A}_R \in \mathbb{R}^{I_1 \times \ldots \times I_N}$ that minimizes

$$f(\mathcal{A}_R) = \frac{1}{2} \|\mathcal{T} - \mathcal{A}_R\|_F^2.$$

FIRST-ORDER OPTIMALITY EQUATIONS

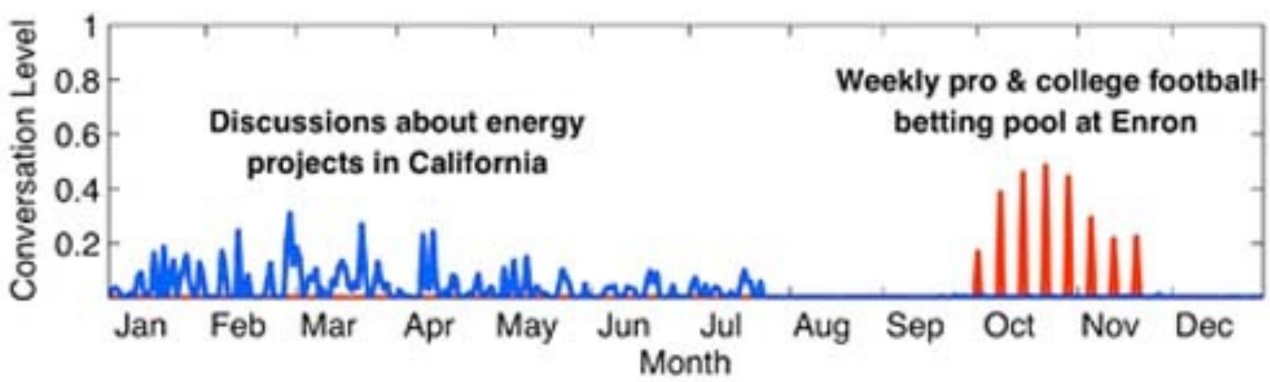
$$\nabla f(\mathcal{A}_R) = \mathbf{g}(\mathcal{A}_R) = 0.$$

(problem is non-convex, multiple (local) minima, solution may not exist (ill-posed), ... ; but smooth, and we assume there is a local minimum)

[de Silva and Lim, 2009]

tensor approximation applications

"Discussion Tracking in Enron Email Using PARAFAC" by Bader, Berry and Browne (2008) (sparse, nonnegative)



‘workhorse’ algorithm: alternating least squares (ALS)

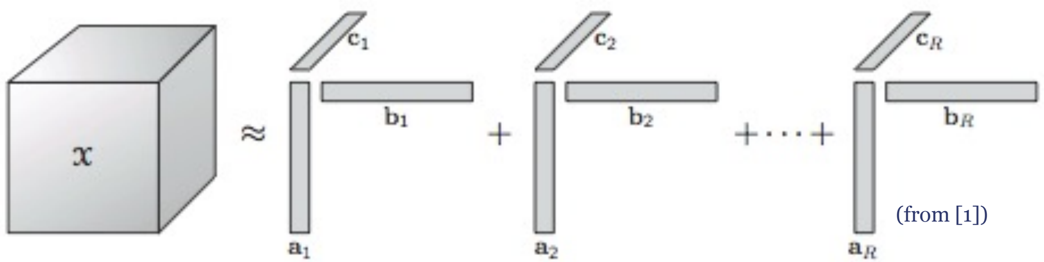
$$f(\mathcal{A}_R) = \frac{1}{2} \left\| \mathcal{T} - \sum_{r=1}^{R} a_r^{(1)} \circ a_r^{(2)} \circ a_r^{(3)} \right\|_F^2$$

(1) freeze all $a_r^{(2)}, a_r^{(3)}$, compute optimal $a_r^{(1)}$ via a least-squares solution (linear, overdetermined)

(2) freeze $a_r^{(1)}, a_r^{(3)}$, compute $a_r^{(2)}$

(3) freeze $a_r^{(1)}, a_r^{(2)}$, compute $a_r^{(3)}$

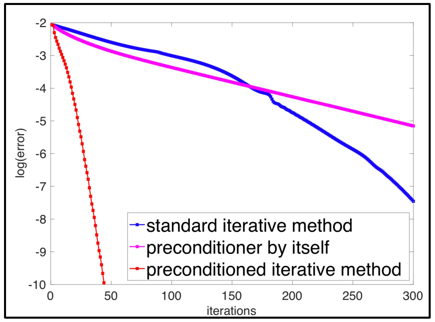
- repeat



(from [1])

alternating least squares (ALS)

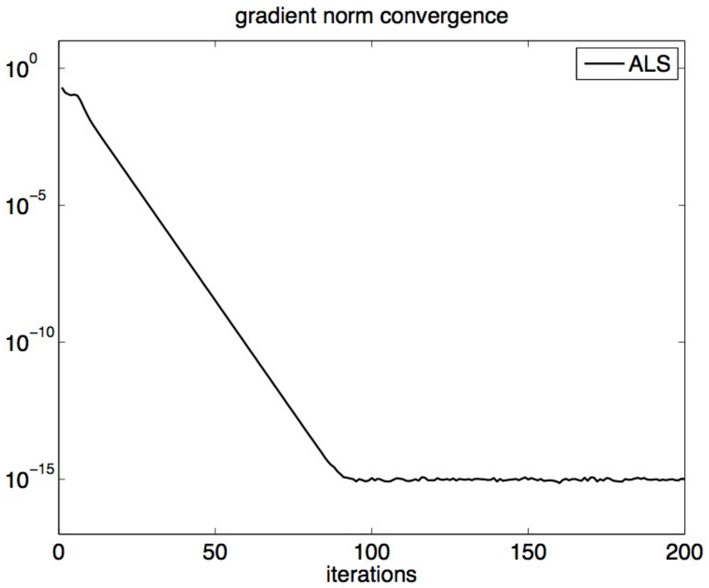
(block nonlinear Gauss-Seidel)

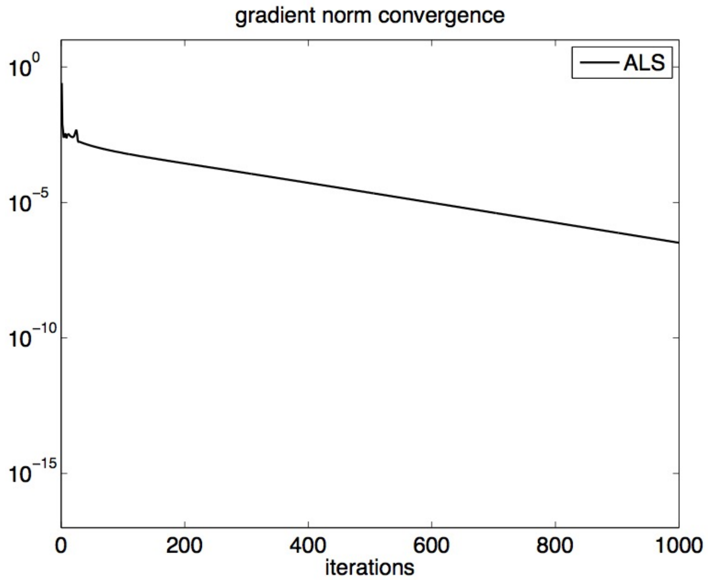


$$f(\mathcal{A}_R) = \frac{1}{2} \left\| \mathcal{T} - \sum_{r=1}^{R} a_r^{(1)} \circ a_r^{(2)} \circ a_r^{(3)} \right\|_F^2$$

fast case

slow case





- can we accelerate ALS using AA (in this nonlinear case), just like SSOR is accelerated by GMRES?

(2) convergence acceleration methods

- Anderson Acceleration: Anderson, D.G.: Iterative procedures for nonlinear integral equations. Journal of the ACM (JACM) **12**(4), 547–560 (1965)

$$x_{k+1} = q(x_k) + \sum_{i=1}^{\min(k,m)} \beta_i^{(k)} (q(x_k) - q(x_{k-i}))$$

- Nonlinear GMRES (NGMRES): T. Washio and C. W. Oosterlee, *Krylov subspace acceleration for nonlinear multigrid schemes*, Electronic Transactions on Numerical Analysis, 6 (1997), pp. 3–1.

$$x_{k+1} = q(x_k) + \sum_{i=0}^{\min(k,m)} \beta_i^{(k)} (q(x_k) - x_{k-i})$$

- both AA and NGMRES reduce to GMRES if $\mathbf{q}(\mathbf{x})$ is linear
- also, other acceleration methods can be used for $\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$: NCG, LBFGS, Nesterov with restart, adaptive algebraic multigrid (not considered here) (De Sterck et al., 2012a, 2012b, 2013, 2015a, 2015b, 2016, 2017, 2020; applied to ALS for tensor decomposition)
- $\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$ can be seen as a nonlinear preconditioner for AA or NGMRES

(3) convergence theory: linear convergence factors

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$$

Definition 2.2 (r-linear convergence). *Let $\{x_k\}$ be any sequence that converges to x^*. Define*

$$\rho_{\{x_k\}} = \limsup_{k\to\infty} \|x_k - x^*\|^{\frac{1}{k}}.$$

We say $\{x_k\}$ converges r-linearly with r-linear convergence factor $\rho_{\{x_k\}}$ if $\rho_{\{x_k\}} \in (0,1)$ and r-superlinearly if $\rho_{\{x_k\}} = 0$. The "r-" prefix stands for "root".

Definition 3.1. *Assume that q is a fixed-point iterative process with limit point x^*. We define the set of iteration sequences that converge to x^* as*

$$C(q, x^*) = \Big\{\{x_k\}_{k=0}^{\infty} | \quad x_{k+1} = q(x_k), \forall k = 0, 1, \cdots, \lim_{k\to\infty} x_k = x^*\Big\},$$

and the worst-case r-linear convergence factor over $C(q, x^*)$ as

$$\rho_{q,x^*} = \sup\Big\{\rho_{\{x_k\}} | \quad \{x_k\} \in C(q, x^*)\Big\}. \tag{3.1}$$

We say that the FP method converges r-linearly if $\rho_{q,x^*} \in (0,1)$.

root-linear convergence theorem for differentiable $\mathbf{q}(\mathbf{x})$

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)$$

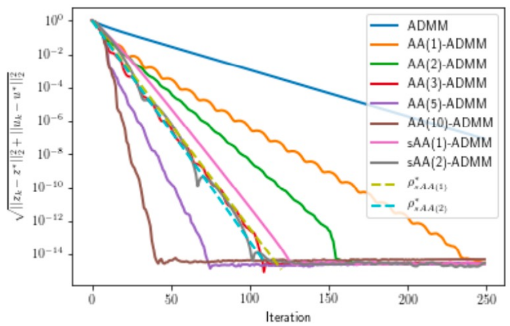
Theorem 3.1. *[4, Chapter 10] [Ostrowski Theorem] Suppose that $q : D \subset \mathbb{R}^n \to \mathbb{R}^n$ has a fixed point x^*, an interior point of D, and is differentiable at x^*. If the spectral radius of $q'(x^*)$ satisfies $0 < \rho(q'(x^*)) < 1$, then the FP method converges r-linearly with $\rho_{q,x^*} = \rho(q'(x^*))$.*

- so for ADMM with differentiable $\mathbf{q}(\mathbf{x})$:

$$\rho_{ADMM,\mathbf{x}^*} = \rho(\mathbf{q}'(\mathbf{x}^*))$$

- same for ALS
- but: the iteration function for AA(m) is not differentiable

AA convergence theory



$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i^{(k)} \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$

$$\{\beta_i^{(k)}\} = \underset{\{\beta_i\}}{\operatorname{argmin}} ||\mathbf{r}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i \left(\mathbf{r}(\mathbf{x}_{k-i}) - \mathbf{r}(\mathbf{x}_{k-i-1})\right)||^2$$

- first convergence proof for AA in 2015: $\|e_k\| \leq \frac{1+c}{1-c} c^k \|e_0\|$

  Toth, A., Kelley, C.: Convergence analysis for Anderson acceleration. SIAM Journal on Numerical Analysis **53**(2), 805–819 (2015)

  - AA-ADMM converges (at least) r-linearly with an r-linear convergence factor that is not worse than the convergence factor of ADMM; proof requires boundedness assumption on the beta coefficients, and $\mathbf{q}'(\mathbf{x}^*)$ Lipschitz
  - convergence improvement results (quantifies convergence gain in each iteration):

    C. Evans, S. Pollock, L. G. Rebholz, and M. Xiao, *A proof that anderson acceleration improves the convergence rate in linearly converging fixed-point methods (but not in those converging quadratically)*, SIAM Journal on Numerical Analysis, 58 (2020), pp. 788–810.

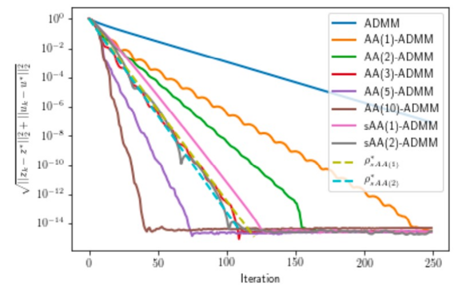
  - no general results exist on improved AA r-linear asymptotic convergence factor

$$\rho_{AA-ADMM,\mathbf{x}^*} = \ ?$$

(4) our contributions: stationary AA(m) and NGMRES(m)

- [1] "On the Asymptotic Linear Convergence Speed of Anderson Acceleration, Nesterov Acceleration, and Nonlinear GMRES", ***De Sterck*** and ***He***, ***SIAM J. Sci. Comp. 2021*** (and ***arXiv:2007.01996***)
  - introduces stationary AA (sAA), and derives optimal convergence theory for sAA
  - ALS for Canonical Tensor Decomposition
- [2] "On the Asymptotic Linear Convergence Speed of Anderson Acceleration Applied to ADMM", ***Wang***, ***He*** and ***De Sterck***, submitted, ***arXiv:2007.02916***
  - ADMM

(optimal) stationary AA (sAA): convergence factor can be analyzed



- AA: $$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i^{(k)} \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$
  - no analysis exists: $\rho_{AA-ADMM,\mathbf{x}^*} = \ ?$
- our sAA: $$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$
  - analyzable: $\rho_{sAA-ADMM,\mathbf{x}^*} = \rho(\mathbf{\Psi}'(\mathbf{x}^*))$
  - e.g., sAA(1): $\mathbf{x}_{k+1} = (1+\beta)\mathbf{q}(\mathbf{x}_k) - \beta\mathbf{q}(\mathbf{x}_{k-1})$

$$\mathbf{X}_{k+1} = \begin{bmatrix} \mathbf{x}_{k+1} \\ \mathbf{x}_k \end{bmatrix} = \begin{bmatrix} (1+\beta)\mathbf{q}(\mathbf{x}_k) - \beta\mathbf{q}(\mathbf{x}_{k-1}) \\ \mathbf{x}_k \end{bmatrix} = \mathbf{\Psi}(\mathbf{X}_k)$$

- sAA with optimal coefficients:
  - given $\mathbf{q}'(\mathbf{x}^*)$, find the optimal β that minimizes $\rho^*_{sAA-ADMM,\mathbf{x}^*} = \rho(\mathbf{\Psi}'(\mathbf{x}^*))$
  - not a practical method, but useful to understand how and to which extent sAA can improve the spectrum of ADMM

theoretical results from [1] on optimal sAA(1) weights

$$\mathbf{x}_{k+1} = (1+\beta)\mathbf{q}(\mathbf{x}_k) - \beta\mathbf{q}(\mathbf{x}_{k-1})$$

$$\mathbf{X}_{k+1} = \begin{bmatrix} \mathbf{x}_{k+1} \\ \mathbf{x}_k \end{bmatrix} = \begin{bmatrix} (1+\beta)\mathbf{q}(\mathbf{x}_k) - \beta\mathbf{q}(\mathbf{x}_{k-1}) \\ \mathbf{x}_k \end{bmatrix} = \mathbf{\Psi}(\mathbf{X}_k)$$

- optimal β :

$$\beta^* = \arg\min_{\beta\in\mathbb{R}} \max\{|\lambda| : \lambda^2 - (1+\beta)\mu\lambda + \beta\mu = 0,\ \mu \in \sigma(\mathbf{q}'(\mathbf{x}^*))\}$$

- first result from [1]: if $\mathbf{q}'(\mathbf{x}^*)$ has real spectrum

Proposition 3 *(Extension of [6, Theorem 3.4].) When* $\sigma(\mathbf{q}'(\mathbf{x}^*)) \subset [0,1)$, *the optimal weight is*

$$\beta^* = \frac{1-\sqrt{1-\sigma_{\max}}}{1+\sqrt{1-\sigma_{\max}}},$$

and the optimal convergence factor is $\rho^*_{sAA(1)} = 1 - \sqrt{1-\sigma_{\max}}$.

theoretical results from [1] on optimal sAA(1) weights

$$\mathbf{x}_{k+1} = (1+\beta)\mathbf{q}(\mathbf{x}_k) - \beta\mathbf{q}(\mathbf{x}_{k-1})$$

$$\mathbf{X}_{k+1} = \begin{bmatrix}\mathbf{x}_{k+1}\\ \mathbf{x}_k\end{bmatrix} = \begin{bmatrix}(1+\beta)\mathbf{q}(\mathbf{x}_k) - \beta\mathbf{q}(\mathbf{x}_{k-1})\\ \mathbf{x}_k\end{bmatrix} = \mathbf{\Psi}(\mathbf{X}_k)$$

- second result from [1]: if $\mathbf{q}'(\mathbf{x}^*)$ has complex spectrum

Proposition 4 *[6] Let the spectral radius of $\mathbf{q}'(\mathbf{x}^*)$ be $\rho^*_{q'}$ and assume $\rho^*_{q'} < 1$. If there exists a real eigenvalue μ of $\mathbf{q}'(\mathbf{x}^*)$ such that $\rho^*_{q'} = \mu$, then the optimal asymptotic convergence rate of sAA(1), $\rho^*_{sAA(1)}$, is bounded below by*

$$\rho^*_{sAA(1)} \geq 1 - \sqrt{1-\rho^*_{q'}},$$

and if the equality holds,

$$\beta^* = \frac{1-\sqrt{1-\rho^*_{q'}}}{1+\sqrt{1-\rho^*_{q'}}}.$$

results from [1] on optimal sAA(m) weights

- sAA(2) and higher: use numerical optimization to find optimal betas

$$\mathbf{x}_{k+1} = (1 + \beta_1 + \beta_2)\mathbf{q}(x_k) - \beta_1 \mathbf{q}(x_{k-1}) - \beta_2 \mathbf{q}(x_{k-2})$$

$$\{\beta_1^*, \beta_2^*\} = \underset{\beta_1,\beta_2 \in \mathbb{R}}{\arg\min} \max_{\lambda} \{|\lambda| : \lambda^3 - (1+\beta_1+\beta_2)\mu\lambda^2 + \beta_1\mu\lambda + \beta_2\mu = 0,\ \mu \in \sigma(\mathbf{q}'(\mathbf{x}^*))\}$$

(5) application to ADMM

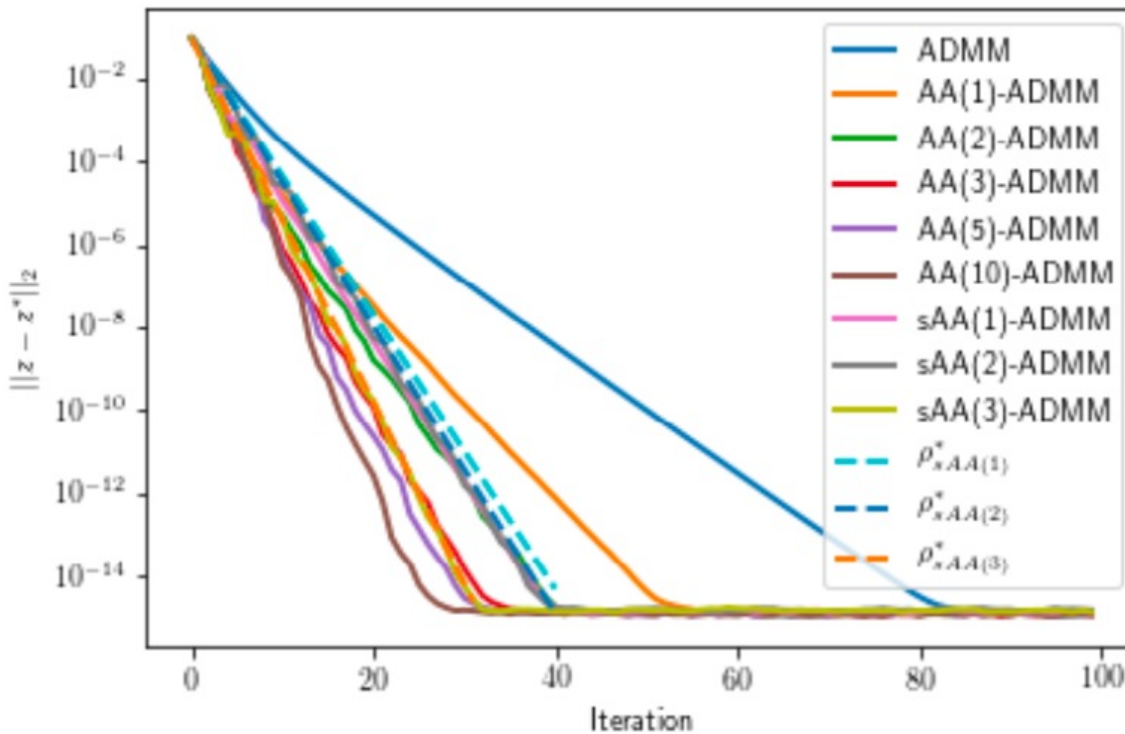
- regularized logistic regression (nonlinear, fully smooth, $\mathbf{q}'(\mathbf{x}^*)$ has real spectrum)

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{m}\sum_{i=1}^{m} \log(1+\exp(-y_i(\mathbf{a}_i^T\mathbf{w}+\mathbf{c}))) + \lambda||\mathbf{z}||_2^2,$$
$$\text{s.t. } \mathbf{x}-\mathbf{z}=0.$$

$$\rho_{ADMM,\mathbf{x}^*} = \rho(\mathbf{q}'(\mathbf{x}^*))$$

$$\rho_{AA-ADMM,\mathbf{x}^*} = \ ?$$

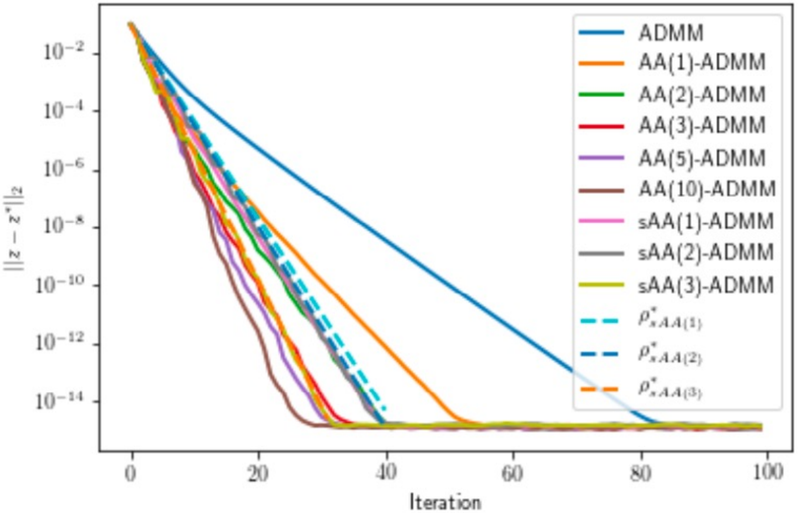
$$\rho^*_{sAA-ADMM,\mathbf{x}^*} = \rho(\mathbf{\Psi}'(\mathbf{x}^*))$$

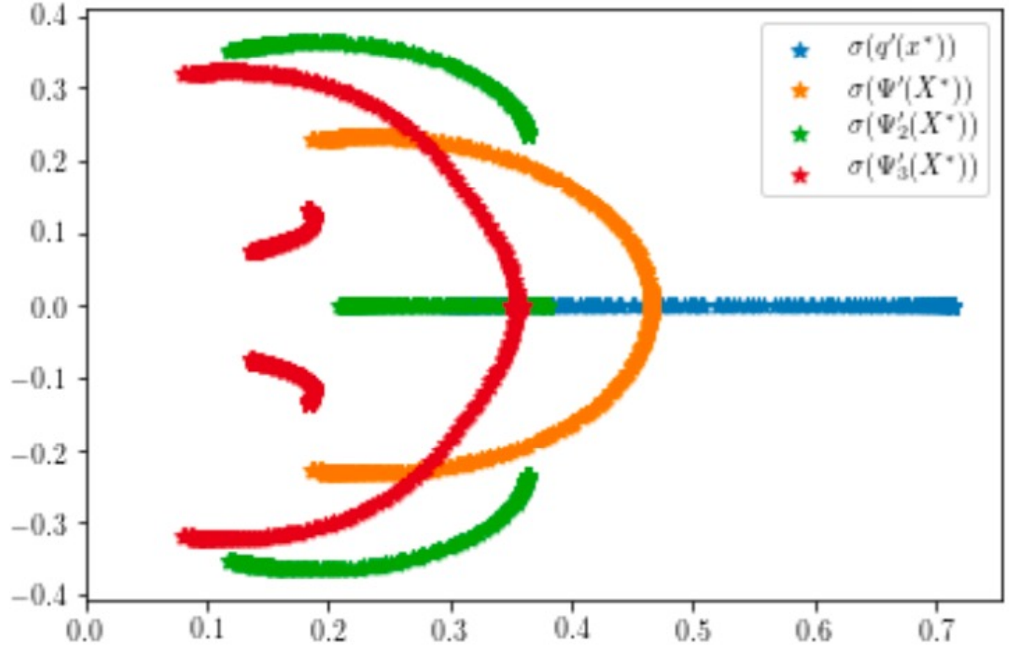


$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i^{(k)} \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$

numerical results – logistic regression





$$\rho_{ADMM,\mathbf{x}^*} = \rho(\mathbf{q}'(\mathbf{x}^*))$$

$$\rho^*_{sAA-ADMM,\mathbf{x}^*} = \rho(\mathbf{\Psi}'(\mathbf{x}^*))$$

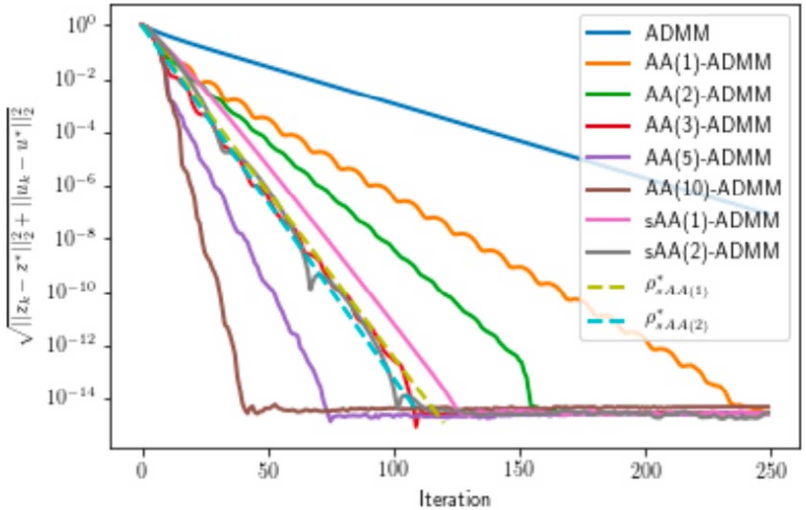
$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$

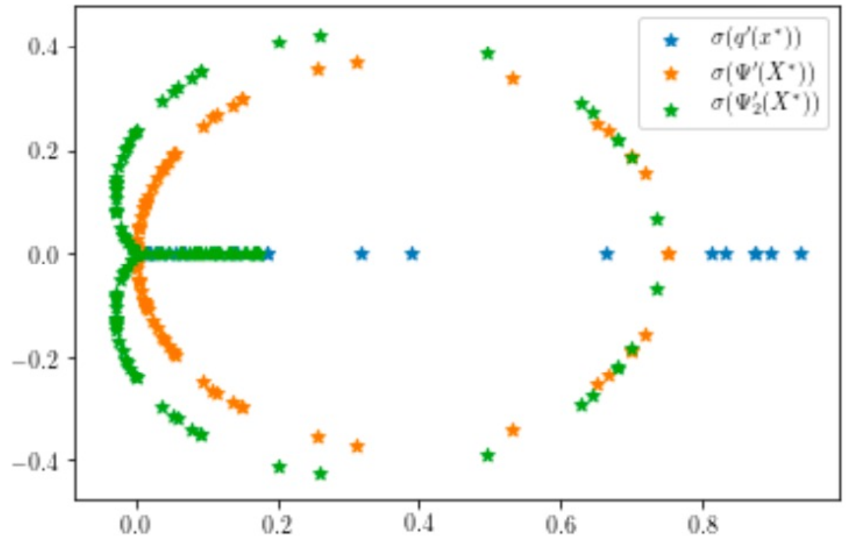
Proposition 1 *Assume $\mu \in \mathbb{R}$. Any complex eigenvalues λ of $\mathbf{\Psi}'(\mathbf{x}^*)$ lie on a circle of radius $\left|\frac{\beta}{1+\beta}\right|$ centered at $(\frac{\beta}{1+\beta}, 0)$ in the complex plane.*

numerical results - LASSO

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}||\mathbf{A}\mathbf{x}-\mathbf{b}||_2^2 + \lambda||\mathbf{z}||_1,$$
$$\text{s.t. } \mathbf{x}-\mathbf{z}=0.$$

- nonlinear, not fully smooth, complex spectrum

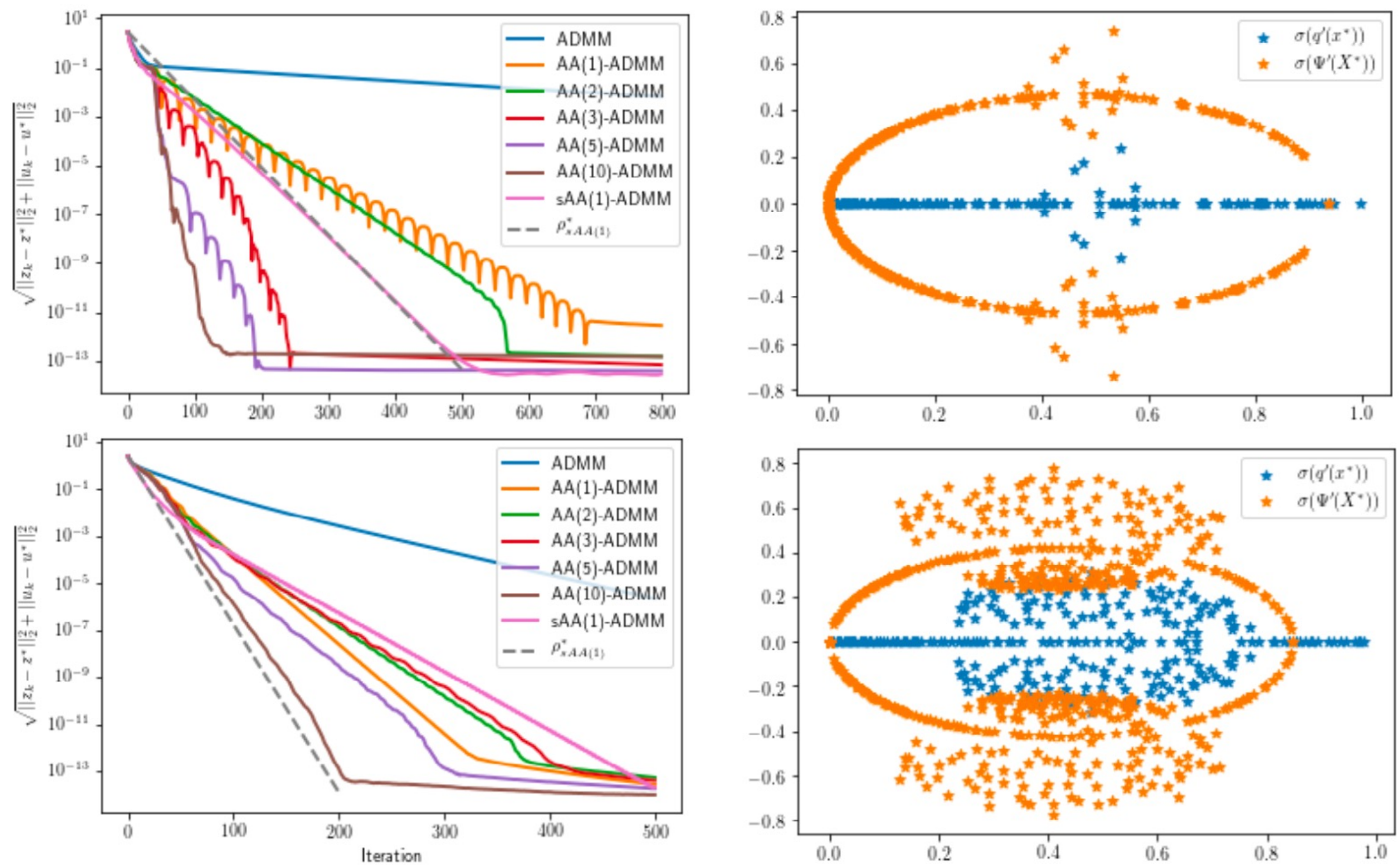




$$\rho_{ADMM,\mathbf{x}^*} = \rho(\mathbf{q}'(\mathbf{x}^*))$$
$$\rho_{AA-ADMM,\mathbf{x}^*} = \ ?$$
$$\rho^*_{sAA-ADMM,\mathbf{x}^*} = \rho(\mathbf{\Psi}'(\mathbf{x}^*))$$

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$

LASSO with increasing density



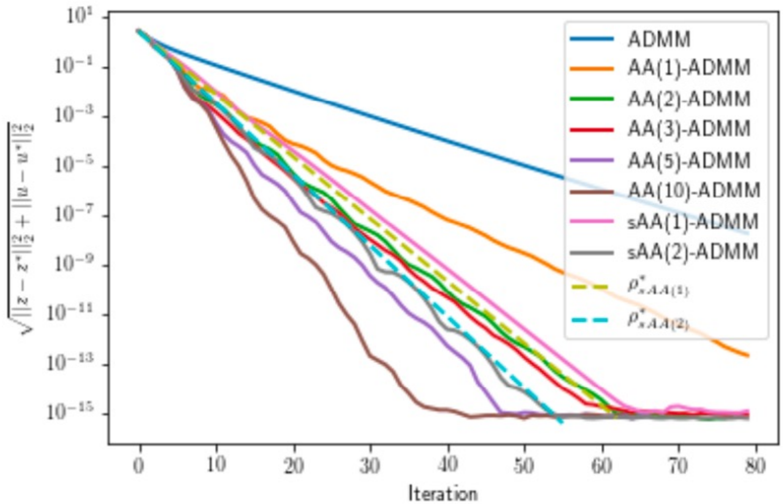
numerical results – nonnegative least squares

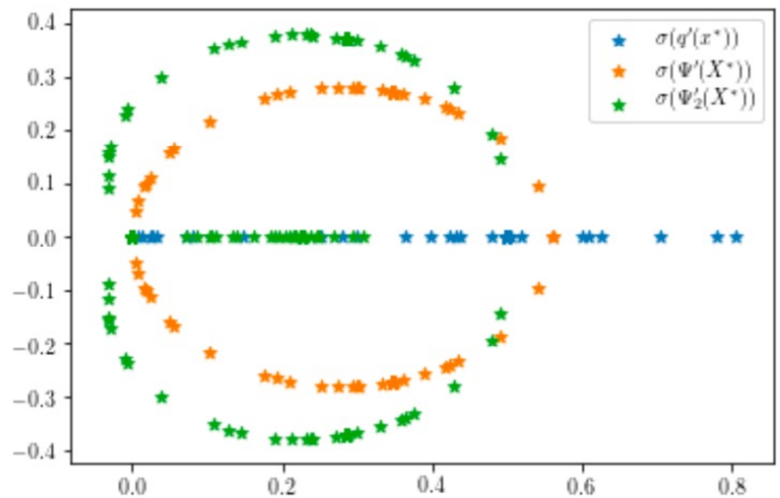
- inequality constraint, real spectrum

$$\min_{\mathbf{x}} ||\mathbf{F}\mathbf{x} - \mathbf{g}||_2^2, \quad \text{s.t.} \quad \mathbf{x} \geq 0.$$

$$\min_{\mathbf{x},\mathbf{z}} \; ||\mathbf{F}\mathbf{x} - \mathbf{g}||_2^2 + \mathcal{I}_{\mathbb{R}_+^n}(\mathbf{z}),$$

$$\text{s.t. } \mathbf{x} - \mathbf{z} = 0,$$





$$\rho_{ADMM,\mathbf{x}^*} = \rho(\mathbf{q}'(\mathbf{x}^*))$$

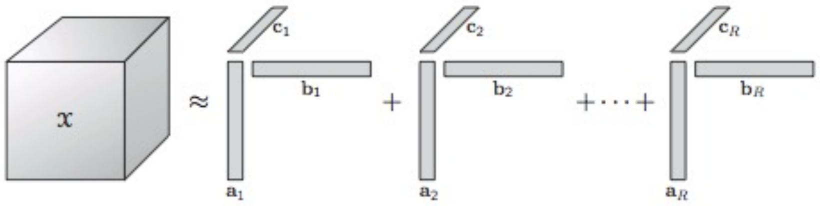
$$\rho_{AA-ADMM,\mathbf{x}^*} = \; ?$$

$$\rho^*_{sAA-ADMM,\mathbf{x}^*} = \rho(\mathbf{\Psi}'(\mathbf{x}^*))$$

$$\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k) + \sum_{i=0}^{m_k-1} \beta_i \left(\mathbf{q}(\mathbf{x}_{k-i}) - \mathbf{q}(\mathbf{x}_{k-i-1})\right)$$

(6) application to canonical tensor decomposition

- tensor decomposition problem



$$[[A^{(1)}, A^{(2)}, \cdots, A^{(N)}]] = \sum_{j=1}^{r} a_j^{(1)} \circ a_j^{(2)} \circ \cdots a_j^{(N)}$$

$$\min f(A^{(1)}, A^{(2)}, \cdots, A^{(N)}) := \frac{1}{2} \left\| \mathcal{Z} - [[A^{(1)}, A^{(2)}, \cdots, A^{(N)}]] \right\|$$

- steepest descent: $x_{k+1} = q_{SD}(x_k) = x_k - \alpha \nabla f(x_k)$

$$q'_{SD}(x) = I - \alpha H(x)$$

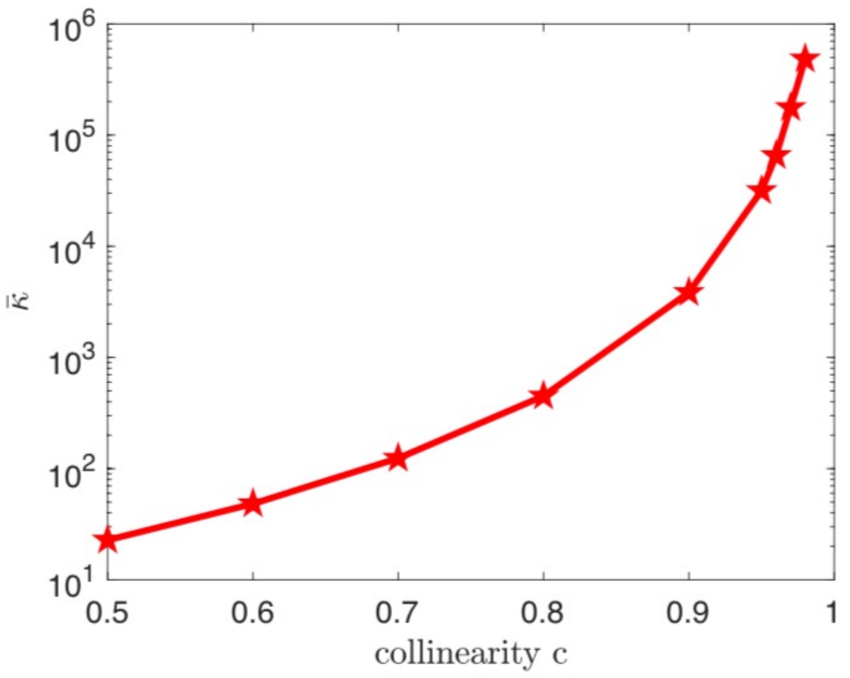
- ALS: $\boxed{\mathbf{x}_{k+1} = \mathbf{q}(\mathbf{x}_k)}$ $\quad q'_{ALS}(x^*) = I - M^{-1}(x^*) H(x^*)$

canonical tensor decomposition

- Hessian has eigenvalues 0 (scaling degeneracy)
- modified Hessian condition number: $\bar{\kappa} = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{L}{\ell}$
- synthetic test problem:

factor matrices have collinearity c
+ noise

ill-conditioned when collinearity close to 1

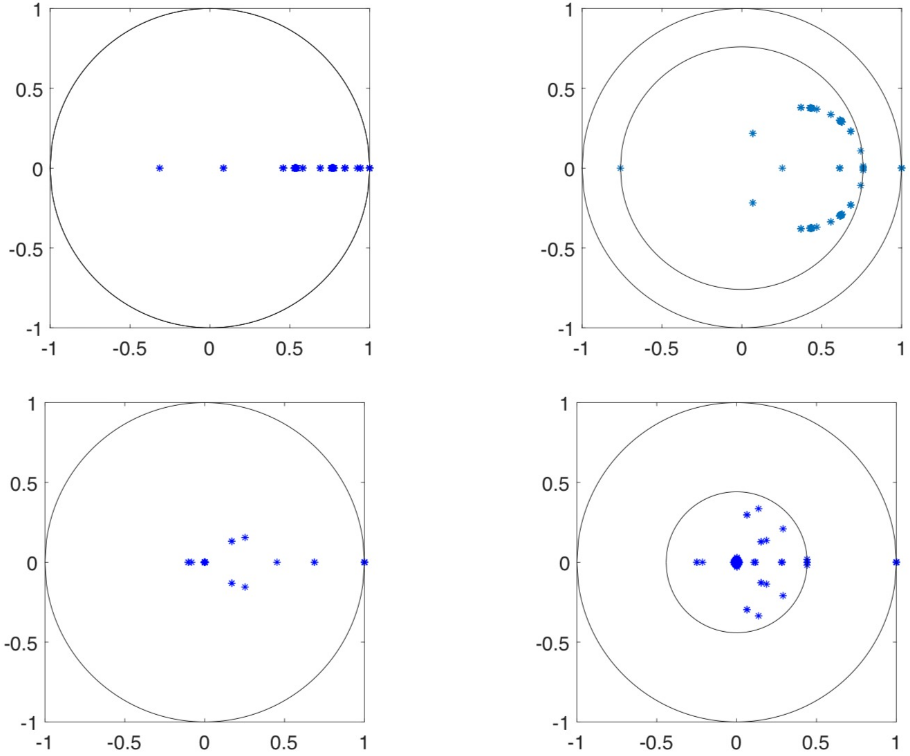


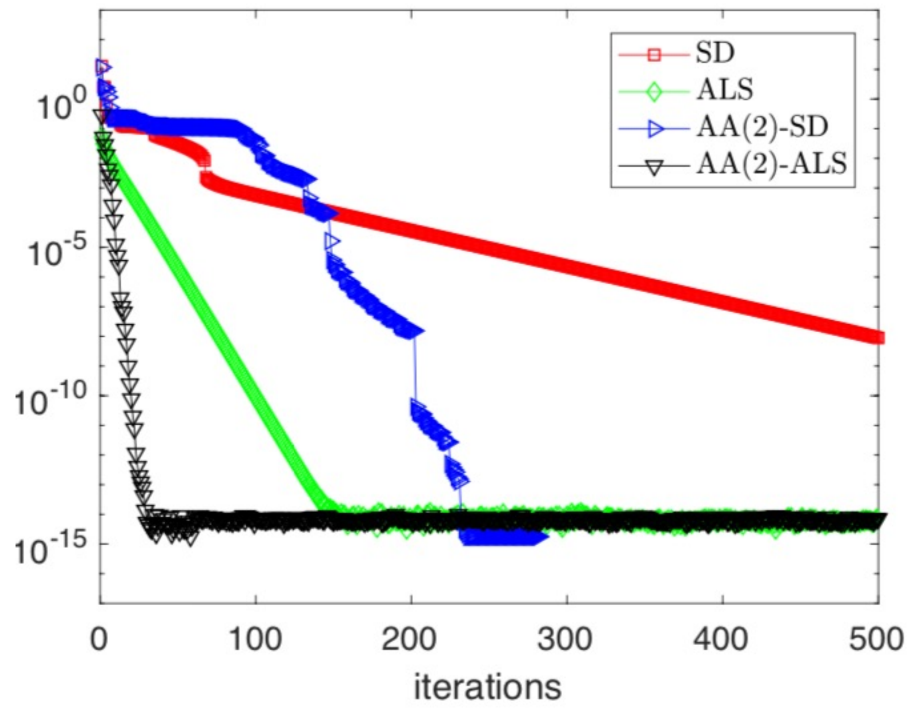
canonical tensor decomposition (c=0.5)

- spectrum of q'(x) and sAA(1)

Ψ'(x), for SD (top)

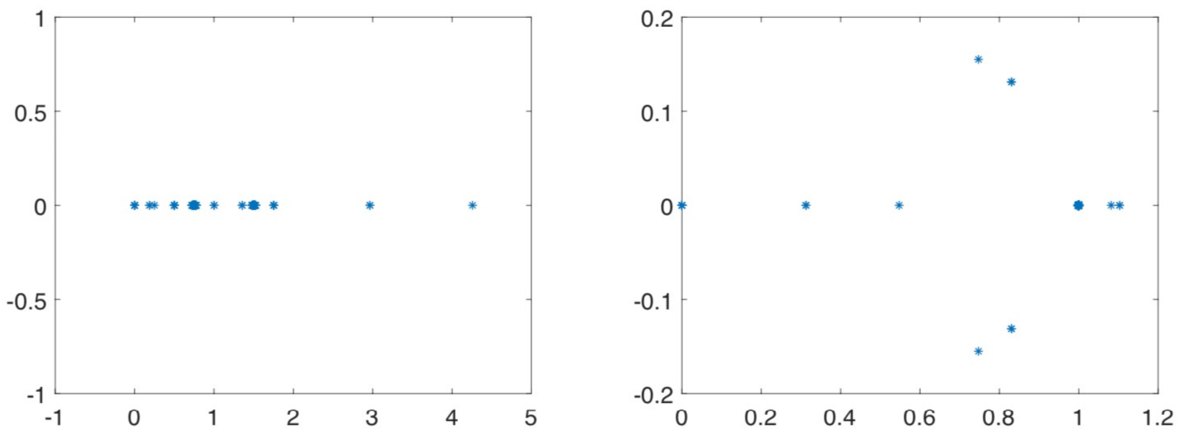
and ALS (bottom)





canonical tensor decomposition

- ALS (right) is a better nonlinear preconditioner than SD (left)



$$\mathbf{q}(\mathbf{x}) = (I - PA)\mathbf{x} + P\mathbf{b}$$

$$q'_{SD}(x) = I - \alpha H(x)$$

$$q'_{ALS}(x^*) = I - M^{-1}(x^*)H(x^*)$$

FIG. 4. *Tensor problem with* $c = 0.5$. *(left) Eigenvalue distribution of* $H(x^*)$*; the (modified) 2-norm condition number* $\bar{\kappa}_2(H(x^*)) = 22.76$. *(right) Eigenvalue distribution of* $M^{-1}(x^*)H(x^*)$*;* $\bar{\kappa}_2(M^{-1}(x^*)H(x^*)) = 7.39$.

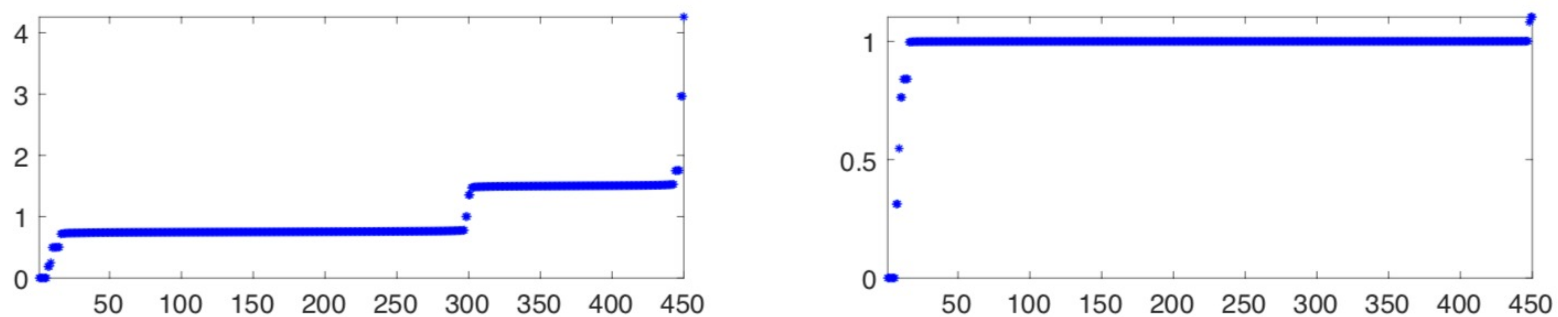


FIG. 5. *Tensor problem with* $c = 0.5$. *(left) Modulus of the eigenvalues of* $H(x^*)$. *(right) Modulus of the eigenvalues of* $M^{-1}(x^*)H(x^*)$.

canonical tensor decomposition

- optimal sAA(1) convergence factor gives good estimate of AA(1) factor

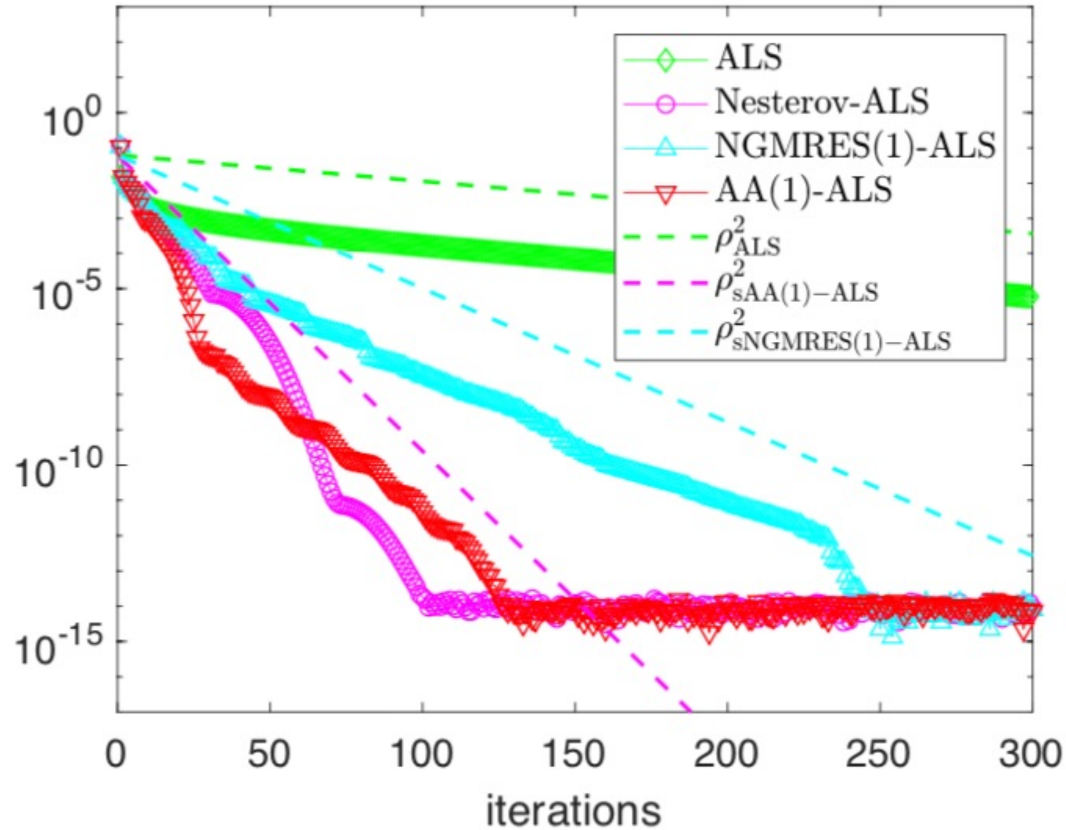


FIG. S.5. *Comparison of the nonstationary AA(1)-ALS, NGMRES(1)-ALS, and Nesterov-ALS methods with theoretical asymptotic convergence factors for optimal stationary methods, for a tensor problem with $c = 0.9$. The vertical axis represents $f(x_k) - f(x^*)$, the convergence towards the minimum value of $f(x)$.*

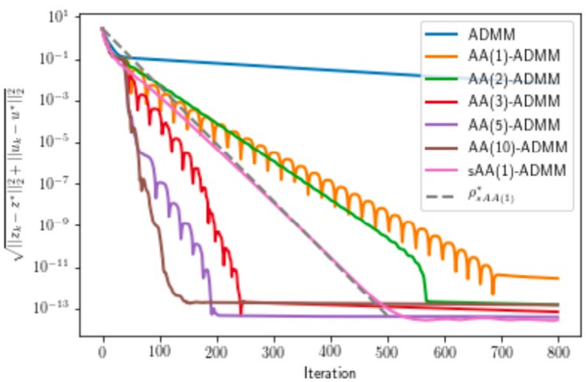
(7) conclusions

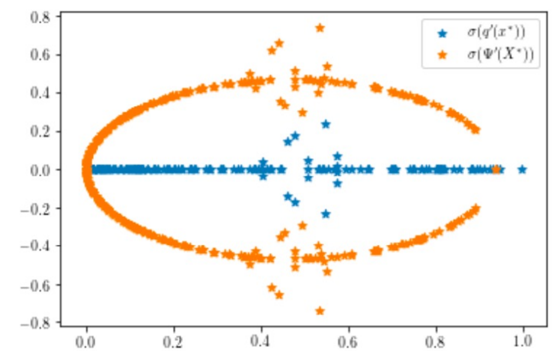
- we have introduced stationary AA and NGMRES, which have computable r-linear convergence factors
- we have derived results on optimal weights for sAA(1) and sNGMRES(1), which allow to understand how and to which extent sAA(1) can improve the spectrum of $\mathbf{q}'(\mathbf{x}^*)$ to obtain faster r-linear converge
- AA or NGMRES with finite window size show similar r-linear convergence improvements as sAA and sNGMRES, but the AA/NGMRES r-linear convergence factors are so far too hard to determine or bound

$$\rho_{ADMM,\mathbf{x}^*} = \rho(\mathbf{q}'(\mathbf{x}^*))$$

$$\rho_{AA-ADMM,\mathbf{x}^*} = \ ?$$

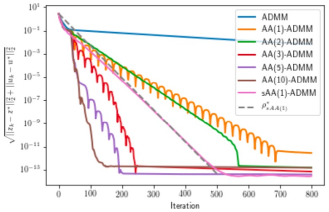
$$\rho^*_{sAA-ADMM,\mathbf{x}^*} = \rho(\mathbf{\Psi}'(\mathbf{x}^*))$$

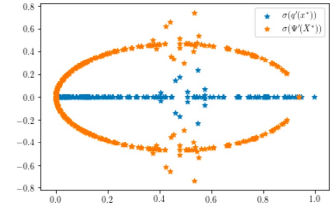




UNIVERSITY OF WATERLOO

conclusions





- the AA/NGMRES r-linear convergence factor (bound) for *finite window size* is expected to depend on the spectral properties of $\mathbf{q}'(\mathbf{x}^*)$ (like for GMRES; e.g. field of values, normality, complex eigenvalues, …)

- for *infinite window size*, from [1]:

CONJECTURE 5.1. *Consider GMRES(∞) applied to linearized fixed-point problem (5.1) with fixed point x^*. If the GMRES residuals satisfy*

$$\frac{\|r_k\|}{\|r_0\|} \le c_1 \rho^k \quad \text{for any } r_0,$$

then the nonlinear residuals of applying NGMRES(∞) and AA(∞) to the nonlinear fixed-point iteration (1.1) associated with (5.1) satisfy

$$\frac{\|r_k\|}{\|r_0\|} \le c_2 \rho^k,$$

provided x_0 is chosen such that the nonlinear methods converge to x^*, and x_0 is chosen sufficiently close to x^*.

$$(I - q'(x^*))\, x = (I - q'(x^*))\, x^*$$

- interesting GMRES results for linear ADMM iteration with infinite window size: Zhang, R.Y., White, J.K.: GMRES-accelerated ADMM for quadratic objectives. SIAM Journal on Optimization **28**(4), 3025–3056 (2018)

Thanks! Questions?

- [1] "On the Asymptotic Linear Convergence Speed of Anderson Acceleration, Nesterov Acceleration, and Nonlinear GMRES", De Sterck and He, SIAM J. Sci. Comp. 2021 (and arXiv:2007.01996)
- [2] "On the Asymptotic Linear Convergence Speed of Anderson Acceleration Applied to ADMM", Wang, He and De Sterck, submitted, arXiv:2007.02916
- Matlab code, acceleration of ALS for tensor problems: https://github.com/hansdesterck/nonlinear-preconditioning-for-optimization
- Python code, acceleration of ADMM: https://github.com/dw-wang/AA-ADMM


UNIVERSITY OF
WATERLOO