



**HAL**  
open science

## GMRES in variable accuracy: a case study in low rank tensor linear systems

Martina Iannacito, Emmanuel Agullo, Olivier Coulaud, Luc Giraud, Gilles Marait, Nick Schenkels

### ► To cite this version:

Martina Iannacito, Emmanuel Agullo, Olivier Coulaud, Luc Giraud, Gilles Marait, et al.. GMRES in variable accuracy: a case study in low rank tensor linear systems. GAMM - Workshop on Applied and Numerical Linear Algebra 2022, Erin Carson; Iveta Hnětynková; Stefano Pozza; Petr Tichý; Miroslav Tůma, Sep 2022, Prague, Czech Republic. hal-03826879v2

**HAL Id: hal-03826879**

**<https://hal.science/hal-03826879v2>**

Submitted on 26 Oct 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# GMRES in variable accuracy: a case study in low rank tensor linear systems

Martina Iannacito

Joint work with: E. Agullo, O. Coulaud, L. Giraud,  
G. Marait and N. Schenkels

GAMM, Prague, 23-09-2022

Concace - Inria joint team with Airbus and Cerfacs

# Table of contents

1. Background on GMRES
2.  $\delta$ -componentwise perturbation
3.  $\delta$ -normwise perturbation
4. Tensor linear systems
5. Parameter dependent tensor linear system

# 1

## Background on GMRES

To solve  $Ax = b$  with initial guess  $x_0 = 0$ , at the  $k$ -th iteration GMRES minimizes the norm of residual

$$\|r_k\| = \min_{x \in \mathcal{K}_k(A, b)} \|Ax - b\|$$

in the space

$$\mathcal{K}_k(A, b) = \text{span}\{b, Ab, \dots, A^{k-1}b\}.$$

From [Drkosova et al. 1995; Paige, Rozložnik, and Strakoš 2006], the recommended stopping criterion is the backward error

$$\eta_{A,b}(x_k) = \min_{\Delta A, \Delta b} \{\tau > 0 : \|\Delta A\| \leq \tau \|A\|, \|\Delta b\| \leq \tau \|b\| \text{ and } (A + \Delta A)x_k = b + \Delta b\}$$

$$\eta_{A,b}(x_k) = \frac{\|Ax_k - b\|}{\|A\|_2 \|x_k\| + \|b\|}$$

Denoting by  $u$  the unit roundoff of the working precision for solving the linear system  $Ax = b$ , then

### Householder-GMRES [Drkosova et al. 1995]

Under technical assumptions [Drkosova et al. 1995, Corollary 4.2], it is shown that the normwise backward error at the last iterate  $x_n$  of Householder-GMRES is such that  $\eta_{A,b}(x_n) = \mathcal{O}(u)$ .

### MGS-GMRES [Paige, Rozložnik, and Strakoš 2006]

If  $\sigma_{\min}(A) \gg n^2 \|A\| u$ , then normwise backward error at the  $k$ -th iterate  $x_k$  of MGS-GMRES is such that  $\eta_{A,b}(x_k) = \mathcal{O}(u)$ , where  $k \in \mathbb{N}$  is the first integer such that  $\kappa_2(\tilde{V}_{k+1}) > 4/3$ .

# 2

$\delta$ -componentwise  
perturbation

## $\delta$ -componentwise storage model

Denoting by  $u$  the unit roundoff of the working precision,

Standard IEEE model [Higham 2002]

$$fl(x \text{ op } y) = (x \text{ op } y)(1 + \varepsilon)$$

with  $|\varepsilon| \leq u$  and  $\text{op} \in \{+, -, \times, \div\}$ .

$\delta$ -componentwise storage

$$\delta\text{-storage}(x) = x(1 + \xi)$$

with  $|\xi| \leq \delta$ .

$\delta$ -representation

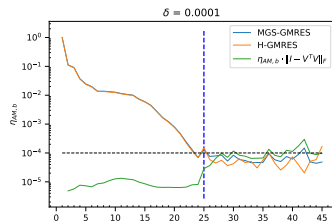
$$fl_{\delta}(x \text{ op } y) = \delta\text{-storage}(fl(x \text{ op } y)) = (x \text{ op } y)(1 + \varepsilon + \xi)$$

with  $|\varepsilon| \leq u$ ,  $|\xi| \leq \delta$  and  $\text{op} \in \{+, -, \times, \div\}$ .

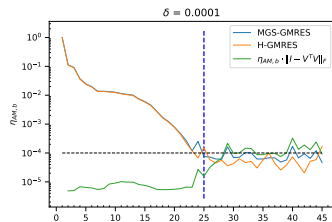


# Numerical experiments $\delta$ -componentwise perturbation

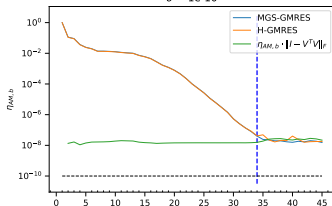
## fp32



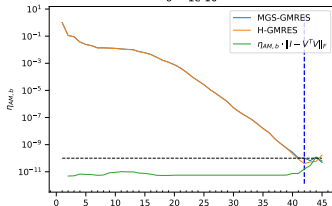
## fp64



$\delta = 1e-10$



$\delta = 1e-10$



The variable accuracy approach reveals the dominating error part.

# 3

$\delta$ -normwise perturbation

Length  $n$  vectors are stored in compressed format with  $\delta$ -normwise representation

### $\delta$ -normwise representation

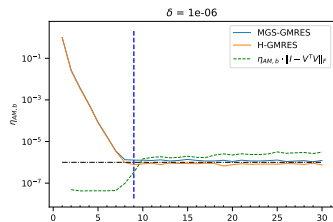
Any vector  $z \in \mathbb{R}^n$  is replaced by  $\bar{z} \in \mathbb{R}^n$  such that

$$\frac{\|z - \bar{z}\|}{\|z\|} \leq \delta.$$

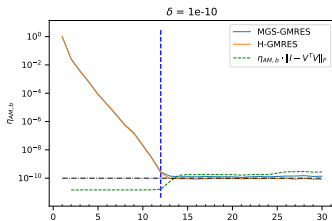
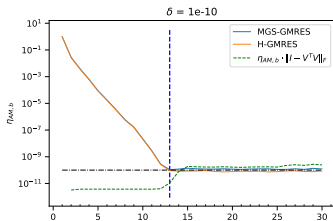
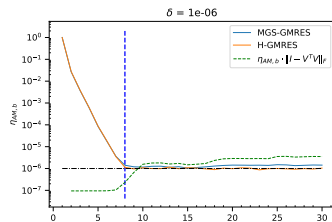
### Remark

The technical details of the theoretical backward stability results in [Drkosova et al. 1995; Paige, Rozloznik, and Strakoš 2006] do not readily apply.

## $\delta$ -componentwise



## $\delta$ -normwise



The  $\delta$ -componentwise and  $\delta$ -normwise perturbation lead to similar results.

# 4

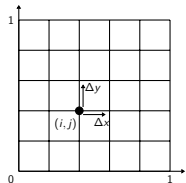
## Tensor linear systems

The problem

$$\begin{cases} \mathcal{L}(u) = f & \text{in } \Omega \\ u = f_0 & \text{in } \partial\Omega \end{cases}$$

for  $\Omega \subseteq \mathbb{R}^{n_1 \times \dots \times n_d}$ .

$\xrightarrow{\text{discretization}}$



$$\mathbf{Ax} = \mathbf{b}$$

where  $\mathbf{A} : \mathbb{R}^{n_1 \times \dots \times n_d} \rightarrow \mathbb{R}^{n_1 \times \dots \times n_d}$  is a multilinear operator and  $\mathbf{b} \in \mathbb{R}^{n_1 \times \dots \times n_d}$  a tensor.

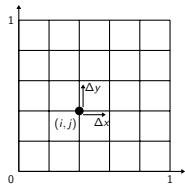
- Curse of dimensionality
- TT-formalism [Oseledets 2011]

The problem

$$\begin{cases} \mathcal{L}(u) = f & \text{in } \Omega \\ u = f_0 & \text{in } \partial\Omega \end{cases}$$

for  $\Omega \subseteq \mathbb{R}^{n_1 \times \dots \times n_d}$ .

→  
discretization



$$\mathbf{Ax} = \mathbf{b}$$

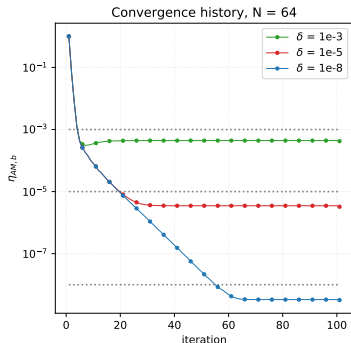
where  $\mathbf{A} : \mathbb{R}^{n_1 \times \dots \times n_d} \rightarrow \mathbb{R}^{n_1 \times \dots \times n_d}$  is a multilinear operator and  $\mathbf{b} \in \mathbb{R}^{n_1 \times \dots \times n_d}$  a tensor.

- Curse of dimensionality
- Memory cost growing with iterations
- TT-formalism [Oseledets 2011]
- TT-rounding [Oseledets 2011]

## TT-representation at accuracy $\delta$

For every TT-vector  $\mathbf{z} \in \mathbb{R}^{n_1 \times \dots \times n_d}$  we compress it getting  $\bar{\mathbf{z}} \in \mathbb{R}^{n_1 \times \dots \times n_d}$  such that

$$\frac{\|\mathbf{z} - \bar{\mathbf{z}}\|}{\|\mathbf{z}\|} \leq \delta.$$



## Convection-Diffusion problem

$$\begin{cases} -\Delta \mathbf{u} + \mathbf{v} \cdot \nabla \mathbf{u} & = 0 \\ \mathbf{u}|_{\{y=1\}} & = 1 \end{cases} \quad \text{in} \quad \Omega = [-1, 1]^3$$



# 5

Parameter dependent  
tensor linear system

Let the parametric convection-diffusion problem be

$$\begin{cases} -\alpha \Delta \mathbf{u} + \mathbf{v} \cdot \nabla \mathbf{u} & = 0 \\ \mathbf{u}|_{\{y=1\}} & = 1 \end{cases} \quad \text{in } \Omega = [-1, 1]^d \text{ and } \alpha \in [1, 10]$$

Solve independently  $p$  tensor linear systems of order  $d$

$$\mathbf{A}_{\alpha_i} \mathbf{x}_{\alpha_i} = \mathbf{b}_{\alpha_i}$$

$$\forall \alpha_i \in \{\alpha_1, \dots, \alpha_p\}.$$

Solve once the tensor linear system  $\mathbf{A} \mathbf{x} = \mathbf{b}$  of order  $(d + 1)$

$$\begin{bmatrix} \mathbf{A}_{\alpha_1} & & \\ & \ddots & \\ & & \mathbf{A}_{\alpha_p} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{[\alpha_1]} \\ \vdots \\ \mathbf{x}^{[\alpha_p]} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_{\alpha_1} \\ \vdots \\ \mathbf{b}_{\alpha_i} \end{bmatrix}$$

What is the numerical quality of  $\mathbf{x}^{[\alpha_i]}$  slice of  $\mathbf{x}$  solution of the  $(d + 1)$  tensor linear system and by construction the solution of the problem  $\mathbf{A}_{\alpha_i} \mathbf{x}_{\alpha_i} = \mathbf{b}_{\alpha_i}$ ?

Let the backward error of the  $(d + 1)$  system  $\mathbf{Ax} = \mathbf{b}$  be

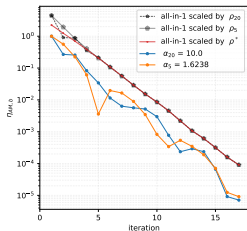
$$\eta_{\mathbf{A},\mathbf{b}}(\mathbf{x}) = \frac{\|\mathbf{b} - \mathbf{Ax}\|}{\|\mathbf{A}\|_2\|\mathbf{x}\| + \|\mathbf{b}\|}.$$

Let  $\mathbf{A}_i, \mathbf{b}_i$  be the multilinear operator and the right-hand side of the parametric problem for the parameter value  $\alpha_i$ , then define

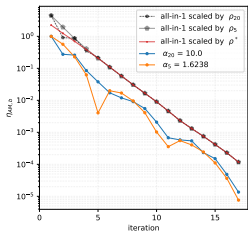
$$\eta_{\mathbf{A}_i,\mathbf{b}_i}(\mathbf{x}^{[i]}) = \frac{\|\mathbf{b}_i - \mathbf{A}_i\mathbf{x}^{[i]}\|}{\|\mathbf{A}_i\|_2\|\mathbf{x}^{[i]}\| + \|\mathbf{b}_i\|}$$

with  $\mathbf{x}^{[i]}$  denotes the  $i$ -th slice of  $\mathbf{x}$  on the parameter mode for  $i \in \{1, \dots, p\}$ . Then we prove that

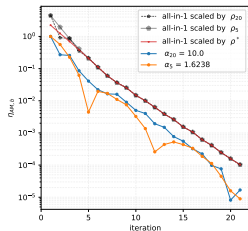
$$\eta_{\mathbf{A},\mathbf{b}}(\mathbf{x}) \rho_i(\mathbf{x}) \geq \eta_{\mathbf{A}_i,\mathbf{b}_i}(\mathbf{x}^{[i]}) \quad \text{where} \quad \rho_i(\mathbf{x}) = \frac{\|\mathbf{A}\|_2\|\mathbf{x}\| + \sqrt{p}}{\|\mathbf{A}_i\mathbf{x}^{[i]}\| + 1}.$$



(I)  $n = 63$



(J)  $n = 127$



(K)  $n = 255$

Figure: 4-d Parametric convection diffusion  $\eta_{AM,b}$  bound using  $\delta = \varepsilon = 10^{-5}$  and  $p = 20$  uniformly logarithmically distributed parameter values  $\alpha_i \in [1, 10]$ .

- Backward stability of GMRES holds numerically [Inria RR-9483] when data are stored with
  - > componentwise perturbation
  - > normwise perturbationboth have practical implementations, for example SZ lossy compressor [S. Di, F. Cappello at Argonne National Lab]
- GMRES in Tensor Train format with TT-rounding seems to be backward stable [Inria RR-9484]
- Special parameter dependent problems of order  $d$  can be solved at once through an order  $(d + 1)$  problem, guaranteeing backward stable bounds linking the  $(d + 1)$  and  $d$  solutions [Inria RR-9484]

Thanks for the attention.

Questions?

If  $\mathbf{M}$  is a preconditioner, the previous stopping criterion gets

$$\eta_{AM,b}(\mathbf{t}_k) = \frac{\|\mathbf{AMt}_k - \mathbf{b}\|}{\|\mathbf{AM}\|_2 \|\mathbf{t}_k\| + \|\mathbf{b}\|} \quad \text{and} \quad \mathbf{x}_k = \mathbf{Mt}_k.$$

Another possible one based just on the right-hand side is

$$\begin{aligned} \eta_b(\mathbf{x}_k) &= \min_{\Delta A, \Delta b} \{ \tau > 0 : \|\Delta b\| \leq \tau \|b\| \text{ and } A\mathbf{x}_k = b + \Delta b \} \\ &= \frac{\|\mathbf{Ax}_k - \mathbf{b}\|}{\|\mathbf{b}\|}. \end{aligned}$$