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

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

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

Inria

1

Background on GMRES

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

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

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

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

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

Backward stability theorem

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, Miroslav, 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

δ-componentwise perturbation

δ-componentwise storage model

Denoting by u the unit roundoff of the working precision,

Standard IEEE model [Higham 2002]

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

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

δ-componentwise storage

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

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

δ-representation

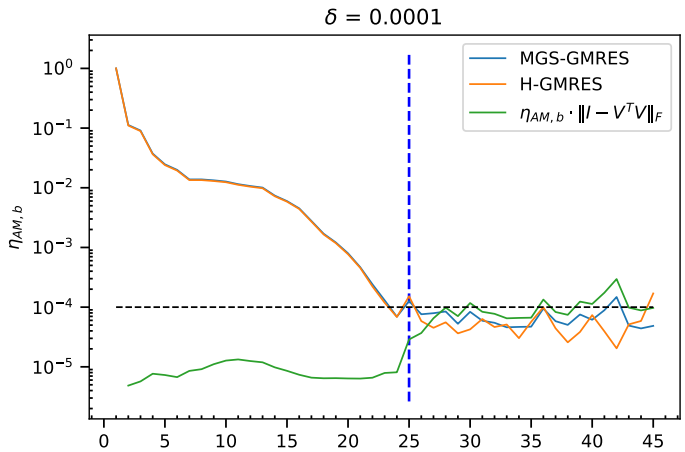
$$fl_\delta(x \operatorname{op} y) = \delta\text{-storage}\big(fl(x \operatorname{op} y)\big) = (x \operatorname{op} y)(1+\varepsilon+\xi)$$

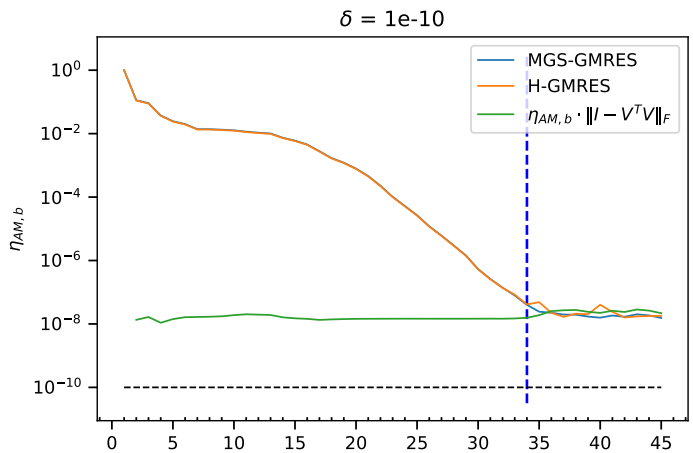
with $|\varepsilon| \leq u$, $|\xi| \leq \delta$ and $\operatorname{op} \in \{+, -, \times, \div\}$.

Inria

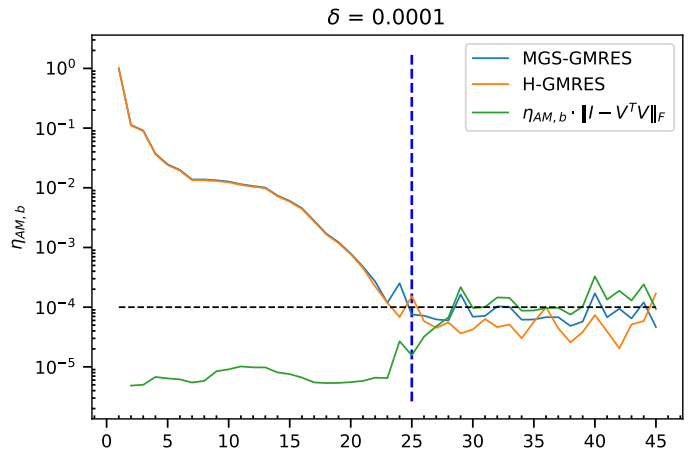
Numerical experiments δ-componentwise perturbation

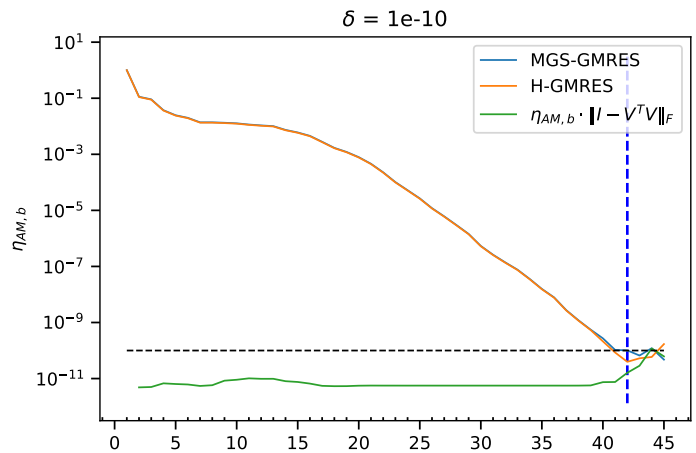
fp32





fp64





The variable accuracy approach reveals the dominating error part.

3

δ-normwise perturbation

δ-normwise storage model

Length n vectors are stored in compressed format with δ-normwise representation

δ-normwise representation

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

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

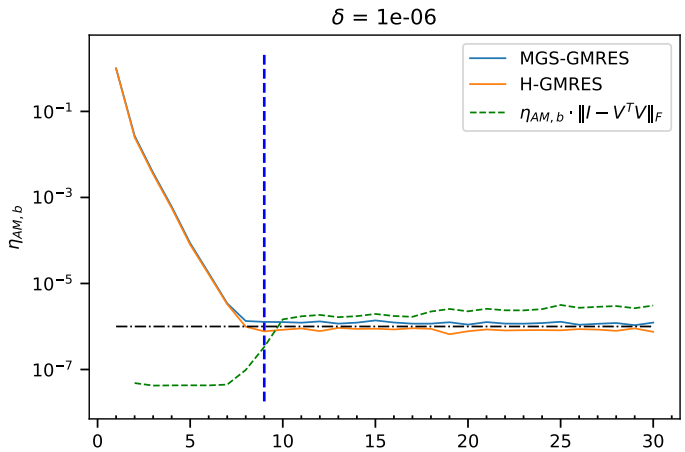
Remark

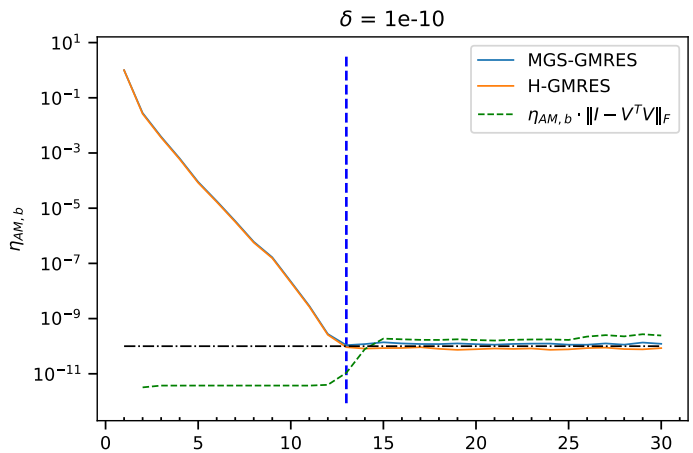
The technical details of the theoretical backward stability results in [Drkosova et al. 1995; Paige, Miroslav, and Strakoš 2006] do not readily apply.

Inria

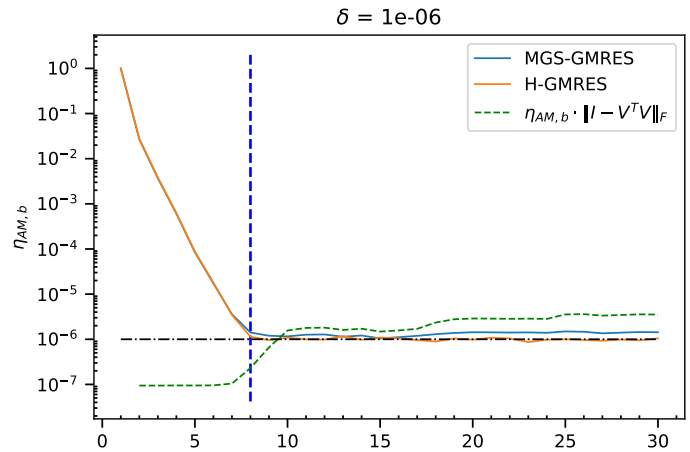
Numerical experiments comparison

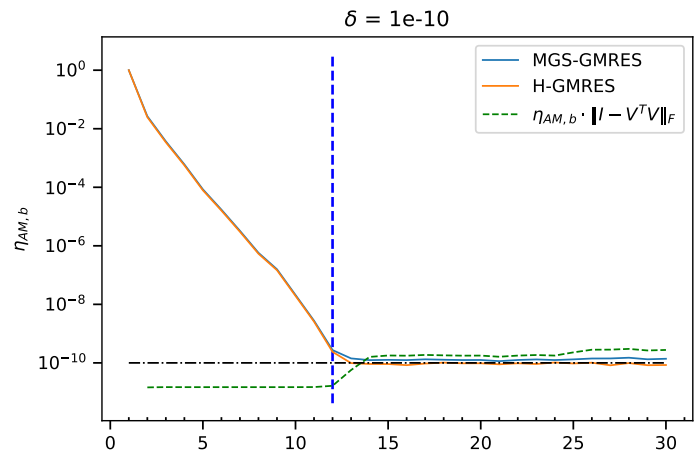
δ-componentwise





δ-normwise





The δ-componentwise and δ-normwise perturbation lead to similar results.

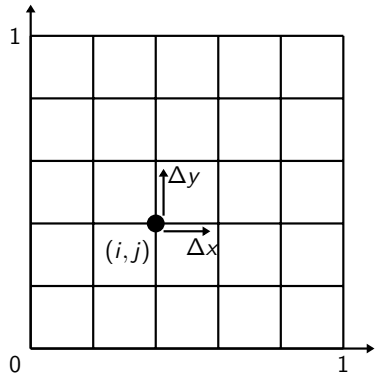
4

Tensor linear systems

The problem

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

for $\Omega \subseteq \mathbb{R}^{n_1 \times \cdots \times n_d}$. $\xrightarrow[\text{discretization}]{}$



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

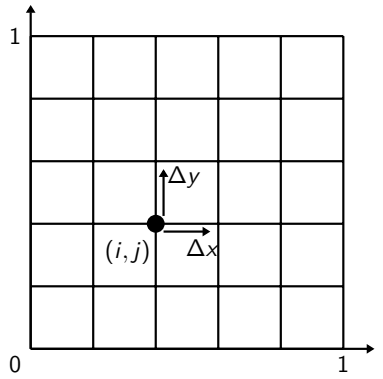
where $\mathbf{A} : \mathbb{R}^{n_1 \times \cdots \times n_d} \to \mathbb{R}^{n_1 \times \cdots \times n_d}$ is a multilinear operator and $\mathbf{b} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ a tensor.

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

The problem

$$\begin{cases} \mathcal{L}(u) & = f \quad \text{in} \quad \Omega \\ u & = f_0 \quad \text{in} \quad \partial\Omega \end{cases} \quad \text{for} \quad \Omega \subseteq \mathbb{R}^{n_1 \times \cdots \times n_d}.$$

$$\xrightarrow[\text{discretization}]{}$$



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

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

- Curse of dimensionality
- Memory cost growing with iterations

- TT-formalism [Oseledets 2011]
- TT-rounding [Oseledets 2011]

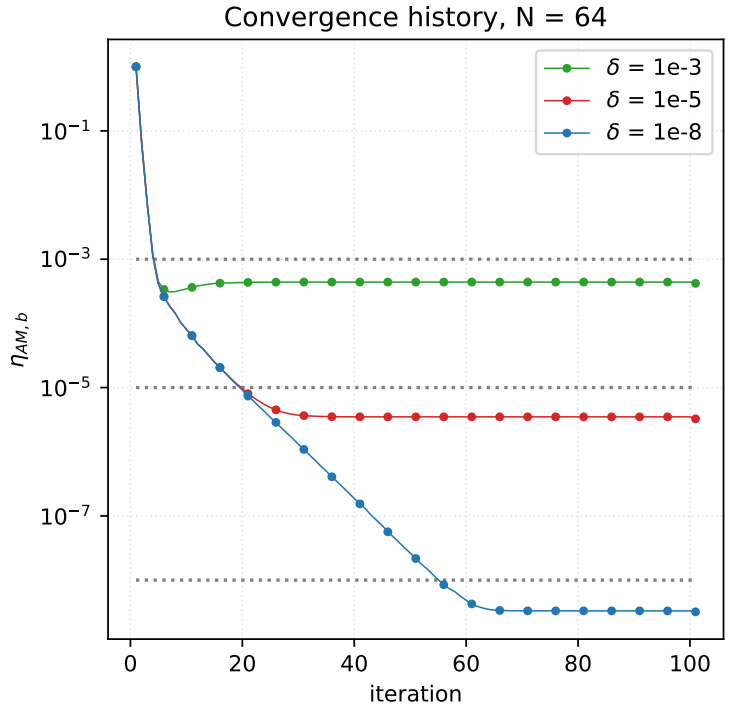
Inria

Robustness of TT-GMRES

TT-representation at accuracy δ

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

Inria

Parameter dependent problem

Let the parametric convection-diffusion problem be

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

Solve independently p tensor linear systems of order d

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

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

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

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

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

Simultaneous vs individual solution bounds

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

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

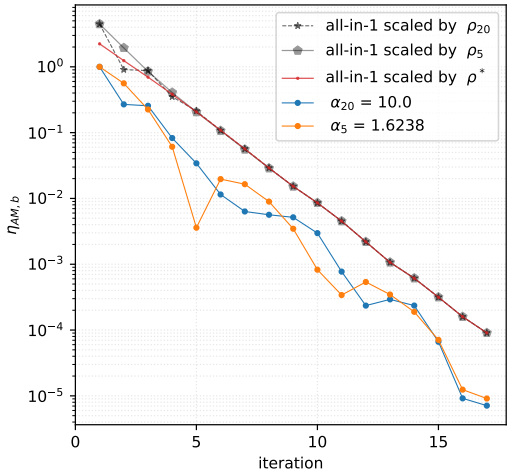
Let $\mathbf{A}_i$, $\mathbf{b}_i$ be the multilinear operator and the right-hand side of the parametric problem for the parameter value α_i, then define

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

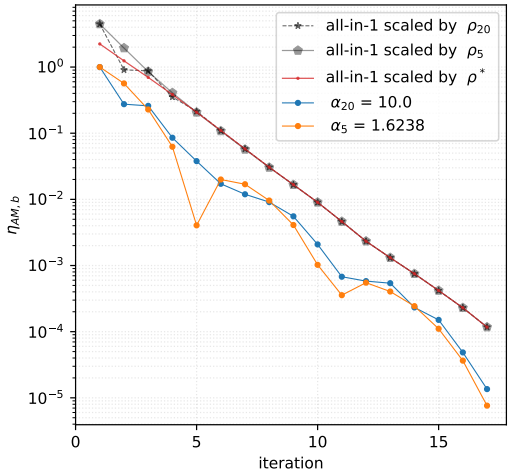
with $\mathbf{x}^{[i]}$ denotes the i-th slice of $\mathbf{x}$ on the parameter mode for $i \in \{1, \dots p\}$. Then we prove that

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

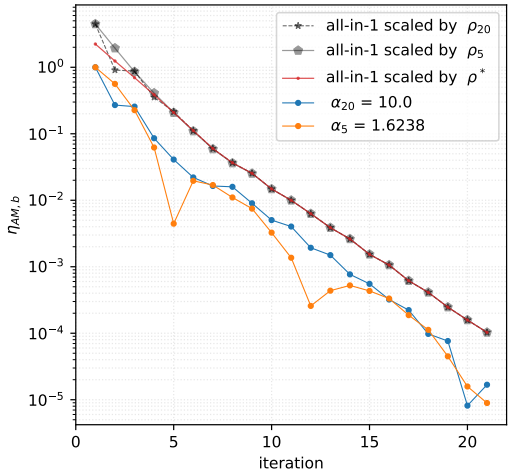
Numerical experiments



(I) $n = 63$



(J) $n = 127$



(K) $n = 255$

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

Take home ideas

- Backward stability of GMRES holds numerically [Inria RR-9483] when data are stored with
  - > componentwise perturbation
  - > normwise perturbation

  both 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?

Alternative backward stable stopping criterion

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

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

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

$$\eta_b(\mathbf{x}_k) = \min_{\Delta A, \Delta b} \{\tau > 0 : ||\Delta b|| \leq \tau ||b|| \text{ and } Ax_k = b + \Delta b\}$$

$$= \frac{||\mathbf{Ax}_k - \mathbf{b}||}{||\mathbf{b}||}.$$