Exploring variable accuracy storage through lossy compression techniques

Application to flexible GMRES

Emmanuel Agullo (Inria), Franck Cappello (ANL), Sheng Di (ANL), Luc Giraud (Inria), Xin Liang (ANL) and Nick Schenkels (Inria)

SIAM LA 21 (virtual conference, May 17 - 21, 2021) - MS63

Outline

Presentation agenda

Accuracy and precision [Higham, 2002]

Accuracy

refers to the absolute or relative error of an approximate quantity.

Precision

is the accuracy with with the basic operations $+$, $-$, $*$, $/$ *are performed.*

Usual implicit assumptions (possible interpretation)

1. All – or most of – the operations are performed with the same precision

2. The absence of a dedicated term for the accuracy with which the numbers are stored may be viewed as if they were stored in the same accuracy as the *precision*: the *precision* can then be interpreted as *both* the accuracy with which the basic operations are *performed* and the numbers are *stored*: [Von Neumann, 1945] (Section 12.2)

   > *the fact that a number requires 32 memory units makes it advisable to subdivide the entire memory in this way [as] it simplifies the organization of the entire memory*

Usual implicit assumptions (possible interpretation)

1. All – or most of – the operations are performed with the same precision, unless stated otherwise such as in a *mixed precision* context: [Wilkinson, 1963] (Section 1.2)

   > *if it is necessary to work to higher precision, [...] we may employ numbers [...] we shall refer to [...] as multiple-precision*

2. The absence of a dedicated term for the accuracy with which the numbers are stored may be viewed as if they were stored in the same accuracy as the *precision*: the *precision* can then be interpreted as *both* the accuracy with which the basic operations are *performed* and the numbers are *stored*: [Von Neumann, 1945] (Section 12.2)

   > *the fact that a number requires 32 memory units makes it advisable to subdivide the entire memory in this way [as] it simplifies the organization of the entire memory*

Usual implicit assumptions (possible interpretation)

1. All – or most of – the operations are performed with the same precision, unless stated otherwise such as in a *mixed precision* context: [Wilkinson, 1963] (Section 1.2)

   > *if it is necessary to work to higher precision, [. . .] we may employ numbers [. . .] we shall refer to [. . .] as multiple-precision*

   Not addressed in this talk.

2. The absence of a dedicated term for the accuracy with which the numbers are stored may be viewed as if they were stored in the same accuracy as the *precision*: the *precision* can then be interpreted as *both* the accuracy with which the basic operations are *performed* and the numbers are *stored*: [Von Neumann, 1945] (Section 12.2)

   > *the fact that a number requires 32 memory units makes it advisable to subdivide the entire memory in this way [as] it simplifies the organization of the entire memory*

Usual implicit assumptions (possible interpretation)

1. All – or most of – the operations are performed with the same precision, unless stated otherwise such as in a *mixed precision* context: [Wilkinson, 1963] (Section 1.2)

   > *if it is necessary to work to higher precision, [. . .] we may employ numbers [. . .] we shall refer to [. . .] as multiple-precision*

   Not addressed in this talk.

2. The absence of a dedicated term for the accuracy with which the numbers are stored may be viewed as if they were stored in the same accuracy as the *precision*: the *precision* can then be interpreted as *both* the accuracy with which the basic operations are *performed* and the numbers are *stored*: [Von Neumann, 1945] (Section 12.2)

   > *the fact that a number requires 32 memory units makes it advisable to subdivide the entire memory in this way [as] it simplifies the organization of the entire memory*

   How about a variable accuracy storage independent from the hardware (*words* in [Wilkinson, 1963] (Section 1.2)) constraints ?

Nothing very new

[Le Verrier, 1840] (contribution to the upcoming Neptune's discovery in 1846)

- Prediction of the existence of a novel planet as well as its location, before its telescopic observation.
- Concerns with *numerical computing* and *significant digits*:

  the coefficients of the equation do not need [...] to be computed with the same approximation [and] one can know which degree of exactness it is necessary to give [to each of them]

This talk

Solution of large sparse linear systems

We consider the solution of linear systems $Ax = b$, with $A \in \mathbb{R}^{n \times n}$ a large and sparse matrix.

Application to the compression of the Z basis in FGMRES [Saad, 1993]

Arnoldi-like equality at step k:

$$AZ_k = V_{k+1}\bar{H}_k, \qquad \text{with} \qquad V_{k+1}^T V_{k+1}.$$

Variable accuracy storage through lossy compression techniques

[Calhoun et al., 2019, Di and Cappello, 2017, Lindstrom, 2014, Lindstrom and Isenburg, 2006, Tao et al., 2017]

Presentation agenda

Generalized minimal residual algorithm (GMRES) (1/2) [Paige et al., 2006, Saad, 2003, Saad and Schultz, 1986]

- Krylov subspace methods remain among the most widely used methods to solve this kind of system and GMRES with a right preconditioner $M \in \mathbb{R}^{n \times n}$ is often the go to method:

$$AM^{-1}u = b \qquad \text{and} \qquad x = M^{-1}u. \tag{1}$$

- Starting from an initial estimate x_0 for $x^\star$, GMRES constructs a series of approximations x_k in Krylov subspaces of increasing size and with decreasing residual norm. More specifically:

$$x_k = \underset{x \in x_{\mathbf{0}} + \mathcal{K}_k(A, r_{\mathbf{0}})}{\arg\min} \|b - Ax\|,$$

with $r_0 = b - Ax_0$ and

$$\mathcal{K}_k(A, r_0) = \text{span}\{r_0, Ar_0, \ldots, A^{k-1}r_0\}$$

the k-dimensional Krylov subspace spanned by A and r_0.

Generalized minimal residual algorithm (GMRES) (2/2)

- In practice, a matrix $V_k = [v_1, \ldots, v_k] \in \mathbb{R}^{n \times k}$ with orthonormal columns and an upper Hessenberg matrix $\bar{H}_k \in \mathbb{R}^{(k+1) \times k}$ are iteratively constructed using the Arnoldi procedure such that span $V_k = \mathcal{K}_k(A, r_0)$ and

$$AV_k = V_{k+1} \bar{H}_k, \qquad \text{with} \qquad V_{k+1}^T V_{k+1}. \tag{2}$$

- This is often referred to as the Arnoldi relation. Consequently, $x_k = x_0 + V_k y_k$ with

$$y_k = \arg\min_{y \in \mathbb{R}^k} \left\| \beta e_1 - \bar{H}_k y \right\|,$$

where $\beta = \|r_0\|$ and $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^{k+1}$.

GMRES with right preconditioning algorithm

input: A, b, x_0, $\texttt{maxit}$, ε, M.
$r_0 = b - Ax_0$, $\beta = \|r_0\|$ and $v_1 = r_0/\beta$
for $k = 1, \dots, \texttt{maxit}$ **do**
  $z = M^{-1} v_k$
  $w = Az$
  for $i = 1, \dots, k$ **do**
    $\bar{H}_{i,k} = v_i^T w$
    $w = w - \bar{H}_{i,k} v_i$
  $\bar{H}_{k+1,k} = \|w\|$
  $v_{k+1} = w / \bar{H}_{k+1,k}$
  $y_k = \arg\min_{y \in \mathbb{R}^k} \left\|\beta e_1 - \bar{H}_k y\right\|$
  $\tilde{r}_k = \beta e_1 - \bar{H}_k y_k$
  if $\|\tilde{r}_k\| < \|b\| \, \varepsilon$ **or** $k = \texttt{maxit}$ **then**
    $x_k = x_0 + M^{-1} V_k y_k$
    $r_k = b - Ax_k$
    if $\|r_k\| < \|b\| \, \varepsilon$ **then**
      break
output: x_k

Flexible GMRES (FGMRES) algorithm I

input: A, b, x_0, $\texttt{maxit}$, ε, M.
$r_0 = b - Ax_0$, $\beta = \|r_0\|$ and $v_1 = r_0/\beta$
for $k = 1, \ldots, \texttt{maxit}$ **do**
    $z_k = M_k^{-1} v_k$
    $w = Az_k$
    for $i = 1, \ldots, k$ **do**
        $H_{i,k} = v_i^T w$
        $w = w - H_{i,k} v_i$
    $H_{k+1,k} = \|w\|$
    $v_{k+1} = w / H_{k+1,k}$
    $y_k = \arg\min_{y \in \mathbb{R}^k} \left\| \beta e_1 - \bar{H}_k y \right\|$
    $\tilde{r}_k = \beta e_1 - \bar{H}_k y_k$
    if $\|\tilde{r}_k\| < \|b\| \varepsilon$ **or** $k = \texttt{maxit}$ **then**
        $x_k = x_0 + Z_k y_k$
        $r_k = b - Ax_k$
        if $\|r_k\| < \|b\| \varepsilon$ **then**

Flexible GMRES (FGMRES) algorithm II

break

output: x_k

FGMRES [Saad, 1993]: remarks

- Main advantage: Increased flexibility for the preconditioner, as now, for example, an iterative method could be used as a preconditioner [Gazzola and Landman, 2019, Giraud et al., 2010, Saad, 1993].
- Main weakness: In contrast to GMRES, Z_k now needs to be stored, because otherwise calculating x_k would require solving all the preconditioning systems an additional time.

Inexact (matrix-vector product) GMRES (1/2)

- Instead of calculating Av, it is actually calculated $(A+E)v$, for some perturbation matrix $E \in \mathbb{R}^{n \times n}$. This idea leads to what is referred to as *inexact Krylov subspace methods* [Bouras and Frayssé, 2005, Giraud et al., 2007, Simoncini and Szyld, 2003, Van Den Eshof and Sleijpen, 2004].
- Again, the Arnoldi relation (2) no longer holds, but it can be shown that the following Arnoldi-like relation holds:

$$AV_k + [E_1 v_1, \ldots, E_k v_k] = V_{k+1} \bar{H}_k. \tag{3}$$

- It turns out that the computed residual in each iteration is given by $\tilde{r}_k = b - \tilde{A}_k x_k$, where $\tilde{A}_k$ is a perturbed version of A and that

$$\tilde{A}_k V_k = V_{k+1} \bar{H}_k, \qquad \text{with} \qquad V_{k+1}^T V_{k+1}.$$

Inexact (matrix-vector product) GMRES (2/2)

- This means that the iterations x_k are in fact members of different Krylov subspaces, each spanned by a different matrix.
- Furthermore, if the size of the perturbations $\|E_k\|$ is bounded in each iteration, it is shown in [Giraud et al., 2007] that the residual gap remains small and that true residual will satisfy the stopping criterion:

Theorem 2.1

Choose $0 < \varepsilon$ and $0 < c < 1$. Define $\varepsilon_c = c\varepsilon$ and $\varepsilon_g = (1-c)\varepsilon$, and assume that in every inexact GMRES iteration k

$$\|E_k\| \leq \frac{c}{n} \sigma_{min}(A) \min\left(1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|} \varepsilon_g\right). \tag{4}$$

Then there exists an $0 < \ell \leq n$ such that $\|\tilde{r}_\ell\| \leq \|b\|\, \varepsilon_c$ and $\|r_\ell\| \leq \|b\|\, \varepsilon$.

Inexact right-preconditioning GMRES

- In [Giraud et al., 2007] it was shown that:

Theorem 2.2

Choose $0 < \varepsilon$ and $0 < c < 1$. Define $\varepsilon_c = c\varepsilon$ and $\varepsilon_g = (1-c)\varepsilon$, and assume that in every GMRES iteration k the right preconditioning system $z = M^{-1}v_k$ is solved with residual p_k. If for all k

$$\|p_k\| \leq \frac{c}{n}\frac{1}{\mathcal{K}\left(AM^{-1}\right)}\min\left(1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|}\varepsilon_g\right), \tag{5}$$

then there exists an $0 < \ell \leq n$ such <that $\|\tilde{r}_\ell\| \leq \|b\|\,\varepsilon_c$ and $\left\|b - AM^{-1}u_\ell\right\| \leq \|b\|\,\varepsilon$.

Proof.

See [Giraud et al., 2007, Theorem 5] for the full proof of this theorem. □

Presentation agenda

FGMRES with inexact right-preconditioning

Theorem 3.1

Choose $0 < \varepsilon$ *and* $0 < c < 1$. *Define* $\varepsilon_c = c\varepsilon$ *and* $\varepsilon_g = (1-c)\varepsilon$, *and assume that in every FGMRES iteration* k *the right preconditioning system* $z_k = M^{-1} v_k$ *is solved with residual* p_k. *If for all* k

$$\|p_k\| \leq \frac{c}{n} \frac{1}{\mathcal{K}\left(AM^{-1}\right)} \min\left(1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|} \varepsilon_g\right), \qquad (6)$$

then there exists an $0 < \ell \leq n$ *such that* $\|\tilde{r}_\ell\| \, \|b\| \, \varepsilon_c$ *and* $\|r_\ell\| \leq \varepsilon \, \|b\|$.

Presentation agenda

Core ideas

- The vectors z_k in FGMRES are the solutions of the preconditioning systems and there are results on preconditioners with lower accuracy [Anzt et al., 2019, Arioli and Duff, 2009, Carson and Higham, 2018, Higham et al., 2019].
- Furthermore, since the z_k can in theory be random – as long as Z_k is of full rank – FGMRES is likely less sensitive to small changes in these vectors.
- In contrast to the mixed precision approaches, however, we will perform all computations in double precision (64 bit), but store the z_k in compressed form after their calculation.
- We note $\tilde{z}_k$ are the vectors containing the decompressed values corresponding to the original z_k.

cFGMRES algorithm I

input: A, b, x_0, $\texttt{maxit}$, ε, M.

$r_0 = b - Ax_0$, $\beta = \|r_0\|$ and $v_1 = r_0/\beta$

for $k = 1, \ldots, \texttt{maxit}$ **do**

  $z_k = M_k^{-1} v_k$

  Compress z_k.

  Retrieve the decompressed vector $\tilde{z}_k$.

  $w = A\tilde{z}_k$

  for $i = 1, \ldots, k$ **do**

    $H_{i,k} = v_i^T w$

    $w = w - H_{i,k} v_i$

  $H_{k+1,k} = \|w\|$

  $v_{k+1} = w / H_{k+1,k}$

  $y_k = \arg\min_{y \in \mathbb{R}^k} \|\beta e_1 - \bar{H}_k y\|$

  $\tilde{r}_k = \beta e_1 - \bar{H}_k y$

  if $\|\tilde{r}_k\| < \|b\| \, \varepsilon$ **or** $k = \texttt{maxit}$ **then**

    Retrieve the decompressed columns of $\tilde{Z}_k = [\tilde{z}_1, \ldots, \tilde{z}_k]$.

cFGMRES algorithm II

$x_k = x_0 + \tilde{Z}_k y_k$
$r_k = b - Ax_k$
if $\|r_k\| < \|b\| \varepsilon$ **then**
  break
output: x

Analysis (framework)

- We write the decompressed values $\tilde{z}_k$ as a perturbed version of the original values z_k:

$$\tilde{z}_k = (I_n + F_k)\, z_k. \tag{7}$$

- Here, $I_n, F_k \in \mathbb{R}^n$ are the identity matrix and a perturbation matrix, respectively. This means that

$$\frac{\|z_k - \tilde{z}_k\|}{\|z_k\|} \leq \zeta_k, \tag{8}$$

with $\zeta_k = \|F_k\|$ the maximum normwise relative compression error in iteration k.

- From a numerical point of view, the only assumption we will make on the compressor is that ζ_k can be controlled by the user.

Analysis (theorem and idea of the proof)

Theorem 4.1

Choose $0 < \varepsilon$ *and* $0 < c < 1$. *Define* $\varepsilon_c = c\varepsilon$ *and* $\varepsilon_g = (1-c)\varepsilon$, *and assume that in every cFGMRES iteration* k *the right preconditioning system* $z_k = M^{-1}v_k = A^{-1}v_k$ *is solved with residual* p_k *and that the maximum normwise relative compression error is given by* $\eta_k > 0$. *If for all* k

$$\|p_k\| + \zeta_k \|A\| \|z_k\| \leq \frac{c}{n} \min\left(1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|}\varepsilon_g\right) \tag{9}$$

then there exists an $0 < \ell \leq n$ *such that* $\|\tilde{r}_\ell\| \|b\| \varepsilon_c$ *and* $\|r_\ell\| \leq \varepsilon \|b\|$.

Proof.

We can interpret the compression as part of the preconditioning and write

$$\tilde{z}_k = (I + F_k)\, M^{-1}\, (v_k - p_k)\,.$$

□

Presentation agenda

Motivation

- Bound (6) from Theorem 3.1 and bound (9) from Theorem 4.1 are both based on results from the theory of inexact Krylov subspace methods, specifically Theorem 2.1.
- In the numerical studies performed in [Bouras and Frayssé, 2005, Simoncini and Szyld, 2003, Van Den Eshof and Sleijpen, 2004] it is, however, shown that this bound is often very restrictive and can be relaxed substantially in many applications.

Strategies (quick overview)

Base strategy

Assuming FGMRES iterations without compression converge, we could ignore the preconditioning error.

Relaxed & double relaxed strategies

- Following [Bouras and Frayssé, 2005, Simoncini and Szyld, 2003, Van Den Eshof and Sleijpen, 2004], we allow larger perturbations in the matrix vector product.
- If the iterations converge, we also have that $\|\tilde{r}_{k-1}\|$ decreases to ε_g, so we can relax this bound a second time (*double relaxed*)

Equal strategy

Theorem 4.1 suggests that it is the *total perturbation* from both the preconditioner and the compression that should be bounded.

Cast 16 & 32 bit (mixed precision *-like*)

Presentation agenda

Numerical set up of the compressor

SZ compressor
[Di and Cappello, 2016, Liang et al., 2018a, Liang et al., 2018b, Tao et al., 2017]

- *prediction based compressor*: meaning that it will try to predict the value of a data point based on the decompressed values of the adjacent data points
- allows one to *control the error* between the original and decompressed data
- used to compress and decompress z

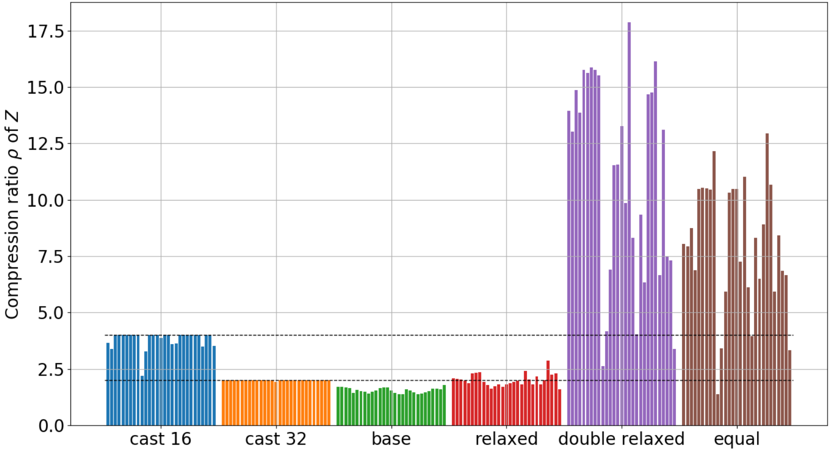
normwise error (SZ interface: $\|z - \tilde{z}\| < \chi$)

- Applied to *base*, *relaxed*, *double relaxed* and *equal* strategies with $\chi_k = \zeta_k \|z_k\|$

pointwise error (SZ interface: $\max_{i=1,\dots,n} \frac{|z[i]-\tilde{z}[i]|}{|z[i]|} < \chi$)

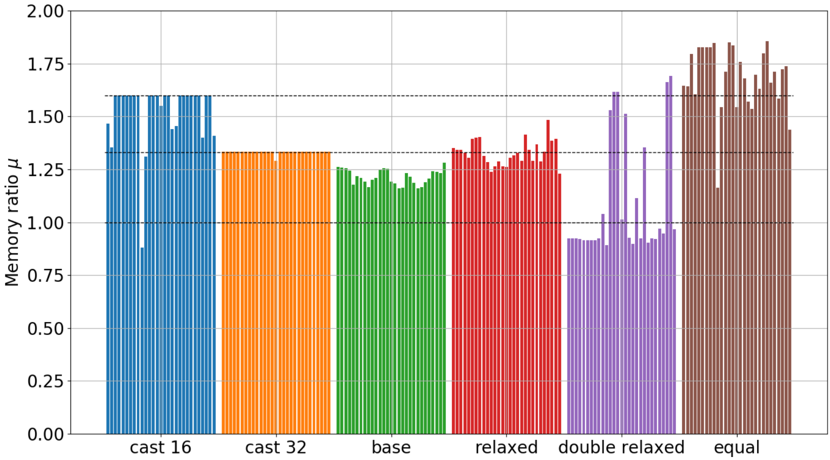
- Applied to *cast32* and *cast64* mixed precision -like strategies

Total compression ratio ρ of Z



- Each bar per strategy corresponds to a different matrix according to its id.
- $\rho = 2$ and 4 represent the cast32 and cast16 respective (numerical) upper bounds.

Total memory ratio μ



- $\mu = 1.33$, 1.6 and 2 represent the cast32, cast16 and cFGMRES respective (numerical) upper bounds.

Presentation agenda

Conclusion

More details in the companion research report [Agullo et al., 2020].

Presentation agenda

References I

Agullo, E., Cappello, F., Di, S., Giraud, L., Liang, X., and Schenkels, N. (2020).
Exploring variable accuracy storage through lossy compression techniques in numerical linear algebra: a first application to flexible GMRES.
Research Report RR-9342, Inria Bordeaux Sud-Ouest.

Anzt, H., Dongarra, J., Flegar, G., Higham, N. J., and Quintana-Ortí, E. S. (2019).
Adaptive precision in block-Jacobi preconditioning for iterative sparse linear system solvers.
Concurrency and Computation: Practice and Experience, 31(6):e4460.

Arioli, M. and Duff, I. S. (2009).
Using FGMRES to obtain backward stability in mixed precision.
Electronic Transactions on Numerical Analysis, 33:31–44.

References II

Bouras, A. and Frayssé, V. (2005).
Inexact matrix-vector products in Krylov methods for solving linear systems: a relaxation strategy.
SIAM Journal on Matrix Analysis and Applications, 26(3):660–678.

Calhoun, J., Cappello, F., Olson, L. N., Snir, M., and Gropp, W. D. (2019).
Exploring the feasibility of lossy compression for PDE simulations.
The International Journal of High Performance Computing Applications, 33(2):397–410.

Carson, E. and Higham, N. J. (2018).
Accelerating the solution of linear systems by iterative refinement in three precisions.
SIAM Journal on Scientific Computing, 40(2):A817–A847.

References III

Di, S. and Cappello, F. (2016).
Fast error-bounded lossy hpc data compression with SZ.
In *2016 IEEE international parallel and distributed processing symposium (ipdps)*, pages 730–739. IEEE.

Di, S. and Cappello, F. (2017).
Optimization of error-bounded lossy compression for hard-to-compress HPC data.
IEEE transactions on parallel and distributed systems, 29(1):129–143.

Gazzola, S. and Landman, M. S. (2019).
Flexible GMRES for total variation regularization.
BIT Numerical Mathematics, 59(3):721–746.

Giraud, L., Gratton, S., and Langou, J. (2007).
Convergence in backward error of relaxed GMRES.
SIAM Journal on Scientific Computing, 29(2):710–728.

References IV

Giraud, L., Gratton, S., Pinel, X., and Vasseur, X. (2010).
Flexible GMRES with deflated restarting.
SIAM Journal on Scientific Computing, 32(4):1858–1878.

Higham, N. J. (2002).
Accuracy and stability of numerical algorithms, volume 80.
SIAM.

Higham, N. J., Pranesh, S., and Zounon, M. (2019).
Squeezing a matrix into half precision, with an application to solving linear systems.
SIAM Journal on Scientific Computing, 41(4):A2536–A2551.

Le Verrier, U. J. (1840).
Mémoire sur les variations séculaires des éléments des orbites: pour les sept planètes principales: Mercure, vénus, la Terre, Mars, Jupiter, Saturne et Uranus.
Journal de mathématiques pures et appliquées, 5:220–254.

References V

Liang, X., Di, S., Tao, D., Chen, Z., and Cappello, F. (2018a).
An efficient transformation scheme for lossy data compression with point-wise relative error bound.
In *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 179–189. IEEE.

Liang, X., Di, S., Tao, D., Li, S., Li, S., Guo, H., Chen, Z., and Cappello, F. (2018b).
Error-controlled lossy compression optimized for high compression ratios of scientific datasets.
In *2018 IEEE International Conference on Big Data (Big Data)*, pages 438–447. IEEE.

Lindstrom, P. (2014).
Fixed-rate compressed floating-point arrays.
IEEE transactions on visualization and computer graphics, 20(12):2674–2683.

References VI

Lindstrom, P. and Isenburg, M. (2006).
Fast and efficient compression of floating-point data.
IEEE transactions on visualization and computer graphics, 12(5):1245–1250.

Paige, C. C., Rozlozník, M., and Strakos, Z. (2006).
Modified Gram-Schmidt (mgs), least squares, and backward stability of MGS-GMRES.
SIAM Journal on Matrix Analysis and Applications, 28(1):264–284.

Saad, Y. (1993).
A flexible inner-outer preconditioned GMRES algorithm.
SIAM Journal on Scientific Computing, 14(2):461–469.

Saad, Y. (2003).
Iterative methods for sparse linear systems, volume 82.
SIAM.

References VII

Saad, Y. and Schultz, M. H. (1986).
GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems.
SIAM Journal on scientific and statistical computing, 7(3):856–869.

Simoncini, V. and Szyld, D. B. (2003).
Theory of inexact Krylov subspace methods and applications to scientific computing.
SIAM Journal on Scientific Computing, 25(2):454–477.

Tao, D., Di, S., Chen, Z., and Cappello, F. (2017).
Significantly improving lossy compression for scientific data sets based on multidimensional prediction and error-controlled quantization.
In *2017 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 1129–1139. IEEE.

References VIII

Van Den Eshof, J. and Sleijpen, G. L. G. (2004).
Inexact Krylov subspace methods for linear systems.
SIAM Journal on Matrix Analysis and Applications, 26(1):125–153.

Von Neumann, J. (1945).
First draft of a report on the EDVAC.

Wilkinson, J. H. (1963).
Rounding errors in algebraic processes.
Prentice-Hall, Inc., Englewood Cliffs, N.J.

Presentation agenda

Matrices (1/2)

	name	n	nnz	$\|\|A\|\|$	iter	η_b	iter
1	atmosmodd	1,270,432	8,814,880	1.92e+05	11	1.38e-11	11
2	atmosmodj	1,270,432	8,814,880	1.92e+05	11	8.43e-11	11
3	atmosmodl	1,489,752	10,319,760	6.20e+05	10	1.37e-11	10
4	atmosmodm	1,489,752	10,319,760	6.39e+06	10	1.15e-11	10
5	cage12	130,228	2,032,536	1.02e+00	8	3.46e-11	8
6	cage13	445,315	7,479,343	1.02e+00	8	5.38e-11	8
7	cage14	1,505,785	27,130,349	1.02e+00	8	5.28e-11	8
8	cage15	5,154,859	99,199,551	1.02e+00	8	9.45e-11	8
9	crashbasis	160,000	1,750,416	6.54e+02	10	3.15e-11	10
10	dc1	116,835	766,396	5.70e+04	139	9.66e-11	11
11	dc2	116,835	766,396	5.84e+04	89	8.80e-11	9
12	dc3	116,835	766,396	6.25e+04	131	9.71e-11	31
13	Goodwin_095	100,037	3,226,066	1.05e+00	245	9.72e-11	120
14	Goodwin_127	178,437	5,778,545	1.05e+00	169	9.66e-11	159
15	hcircuit	105,676	513,072	8.63e+01	215	9.58e-11	30

Matrices (2/2)

	name	n	nnz	$\|\|A\|\|$	iter	η_b	ite
16	language	399,130	1,216,334	2.91e+01	9	3.40e-11	
17	majorbasis	160,000	1,750,416	1.45e+02	10	4.67e-11	1
18	memchip	2,707,524	13,343,948	5.00e+02	68	8.18e-11	
19	ML_Laplace	377,002	27,582,698	2.92e+07	53	8.50e-11	2
20	rajat31	4,690,002	20,316,253	1.25e+04	26	5.26e-11	1
21	ss	1,652,680	34,753,577	6.54e+00	10	5.62e-11	2
22	ss1	205,282	845,089	2.17e+00	7	2.74e-11	
23	stomach	213,360	3,021,648	2.21e+00	10	4.00e-11	1
24	torso2	115,967	1,033,473	8.06e+00	10	2.60e-11	
25	trans5	116,835	749,800	1.13e+04	417	9.56e-11	1
26	Transport	1,602,111	23,487,281	1.00e+00	34	7.55e-11	2
27	vas_stokes_1M	1,090,664	34,767,207	8.85e+00	76	8.57e-11	7
28	vas_stokes_2M	2,146,677	65,129,037	8.19e+00	72	5.77e-11	6
29	xenon2	157,464	3,866,688	5.29e+28	22	7.87e-11	2

Presentation agenda

Base strategy

- As stated before, in practice it is observed that $\|p_k\|$ can be larger than what Theorem 4.1 would suggest. Assuming that the FGMRES iterations without compression converge, we could ignore the preconditioning error and only try to bound the compression error using (9), i.e.,

$$\zeta_k \leq \frac{c}{n\,\|A\|\,\|z_k\|} \min\left(1, \frac{\|b\|}{\|\tilde{r}_{k-1}\|}\varepsilon_g\right)$$

- In our numerical experiment we will take $c = 0.9$.

Relaxed & double relaxed strategies

- The bound used in Theorem 2.1 can be written as

$$\|E_k\| \leq \lambda_k \frac{1}{\|\tilde{r}_{k-1}\|} \varepsilon_g$$

- In [Bouras and Frayssé, 2005, Simoncini and Szyld, 2003, Van Den Eshof and Sleijpen, 2004], it is shown that that setting $\lambda_k = 1$, thus allowing larger perturbations in the matrix vector product, does not negatively impact the convergence in many cases.
- We will therefore do the same with the base compression strategy and relax bound (9) to find

$$\zeta_k \leq \frac{1}{\|A\| \, \|z_k\| \, \|\tilde{r}_{k-1}\|} \varepsilon_g. \tag{10}$$

- If the iterations converge, we also have that $\|\tilde{r}_{k-1}\|$ decreases to ε_g, so we can relax this bound a second time by replacing $\varepsilon_g / \|\tilde{r}_{k-1}\|$ with 1:

$$\zeta_k \leq \frac{1}{\|A\| \, \|z_k\|}. \tag{11}$$

- We will refer to strategy (10) and (11) as the *relaxed* and *double relaxed* strategies, respectively.

Equal strategy

- Assuming that the FGMRES iterations without compression converge, there is a series of preconditioning errors $\|p_k\|$ which do not prevent the algorithm from converging. Instead of using the upper bound from (9), we could relax the base strategy by using $\|p_k\|$ as an upper bound for the maximum normwise relative compression error in each iteration, i.e.,

$$\zeta_k \|z_k\| \|A\| \leq \|p_k\| \;\Leftrightarrow\; \zeta_k \leq \frac{\|p_k\|}{\|z_k\| \|A\|}.$$

- Another way to interpret this strategy is to note that Theorem 4.1 suggests that it is the *total perturbation* from both the preconditioner and the compression that should be bounded. If the compression error in each iteration is less then or equal to the preconditioning error, then

$$\|p_k\| + \eta_k \|A\| \|z_k| \leq 2 \|p_k\|,$$

implying that the order of magnitude of the total perturbation has remained equal to that of the FGMRES iterations without compression – which we assumed converged.

Cast 16 & 32 bit (mixed precision *-like*)

- Due to the large interest in mixed precision arithmetic we will also compare the previous compression strategies with a mixed precision inspired approach: storing the z_k in either 16 bit or 32 bit precision.
- We will, however, perform all calculations in 64 bit, and the decompression step will therefore consist of casting the vector back to 64 bit.
- Additionally, in order to limit over- and underflow errors when casting – especially to 16 bit – we will normalize z_k before casting it and store the norm of the original data as well. After the vector is cast back to 64 bit we multiply it with its original norm in order to retrieve the decompressed vector $\tilde{z}_k$.

Presentation agenda

Individual compression ratio (preliminary note)

Individual compression ratio ρ_k in iteration k

Ratio of saved storage for the z_k stored as $\bar{z}_k$ as

$$\rho_k = \frac{\mathtt{mem}(z_k)}{\mathtt{mem}(\bar{z}_k)} = \frac{\mathtt{mem}(z)}{\mathtt{mem}(\bar{z}_k)}.$$

where:

- $\bar{\cdot}$: compressed data object
- $\mathtt{mem}(\cdot)$: memory used by an object.

Remarks

- Since $\mathtt{mem}(z_k)$ is equal for all k, we will simply write $\mathtt{mem}(z)$.
- Note that the memory used by $\bar{z}_k$ can vary because the compression ratio depends on z_k itself and on the bound for the pointwise relative error – which will vary in each iteration.

Compression ratio (metric 1)

Compression ratio ρ of Z

If FGMRES needs ℓ_{ref} iterations to converge and cFGMRES ℓ iterations then we define the total compression ratio ρ associated with $Z_\ell = [z_1, \ldots, z_\ell]$ as

$$\rho = \frac{\sum_{k=1}^{\ell_{ref}} \mathtt{mem}(z_k)}{\sum_{k=1}^{\ell} \mathtt{mem}(\bar{z}_k)} = \frac{\ell_{ref} \cdot \mathtt{mem}(z)}{\sum_{k=1}^{\ell} \frac{\mathtt{mem}(z)}{\rho_k}} = \frac{\ell_{ref}}{\sum_{k=1}^{\ell} \frac{1}{\rho_k}}. \tag{12}$$

Remarks

- The total compression ratio gives us an easy way to asses the overall efficiency of the compression, taking into account the difference in the number of iterations.
- We might, for example, have a high compression ratio in each iteration, but if we need many extra iterations to converge, we may eventually have $\rho < 1$.

Memory ratio (metric 2)

Memory ratio μ

In order to estimate how much memory we gain with respect to FGMRES we also define the total memory ratio μ that takes into account the storage required for both the v_k and the z_k:

$$\begin{aligned}\mu = \frac{\sum_{k=1}^{\ell_{ref}} \mathtt{mem}(v_k) + \mathtt{mem}(z_k)}{\sum_{k=1}^{\ell} \mathtt{mem}(v_k) + \mathtt{mem}(\bar{z}_k)} &= \frac{\ell_{ref} \cdot (\mathtt{mem}(v) + \mathtt{mem}(z))}{\ell \cdot \mathtt{mem}(v) + \sum_{k=1}^{\ell} \mathtt{mem}(\tilde{z}_k)} \\ &= \frac{2\ell_{ref} \cdot \mathtt{mem}(z)}{\ell \cdot \mathtt{mem}(z) + \sum_{k=1}^{\ell} \frac{\mathtt{mem}(z)}{\rho_k}} \\ &= \frac{2\ell_{ref}}{\ell + \sum_{k=1}^{\ell} \frac{1}{\rho_k}}.\end{aligned} \tag{13}$$

Remark

- Here we use the fact that $\mathtt{mem}(v_k) = \mathtt{mem}(v) = \mathtt{mem}(z)$ for all k.

Extra remarks

- Obviously, higher individual compression ratios ρ_k in each iteration k will lead to a higher total compression ratio ρ and total memory ratio μ. The latter (the memory ratio μ) will, however, penalize extra iterations a lot more than the former since it takes into account the fact that the extra v_k need to be stored as well – without compression. Note that we can write

$$\mu(\rho) = \frac{2\ell_{ref}\rho}{\rho\ell + \ell_{ref}} \quad \Rightarrow \quad \lim_{\rho \to +\infty} \mu(\rho) = 2\frac{\ell_{ref}}{\ell}.$$

- While it is possible that $\ell \leq \ell_{ref}$, we observed in our numerical experiments that the opposite is usually true. This implies that the total memory ratio is bounded by 2, which is not surprising, since even with very high compression rates cFGMRES still needs to store the v_k.
- When the compression is done by casting the z_k to 16 bit and $\ell = \ell_{ref}$, then $\rho = 4$ and $\mu = 1.6$.
- Similarly, for casting to 32 bit we find $\rho = 2$ and $\mu = 4/3 = 1.33$.

Presentation agenda

Acknowledgments

Experiments were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine.