Fast and Accurate Randomized Algorithms for Linear Systems and Eigenvalue Problems

Yuji Nakatsukasa Joel A. Tropp

SIAM J. Matrix Analysis & Applications
Vol. 45, Iss. 2 (2024)
DOI: 10.1137/23M1565413

Motivation and Background

- In scientific computing and machine learning, solving large-scale linear systems $\mathbf{Ax} = \mathbf{f}$ and eigenvalue problems $\mathbf{Ax} = \lambda\mathbf{x}$ is fundamental.
- Classical algorithms such as GMRES (for linear systems) and Rayleigh–Ritz / eigs (for eigenvalue problems) are accurate but costly for large n.
- **Randomized algorithms** (e.g., sketching) enable fast dimension reduction, making traditional solvers scalable.

Motivation and Background

- In scientific computing and machine learning, solving large-scale linear systems $\mathbf{Ax} = \mathbf{f}$ and eigenvalue problems $\mathbf{Ax} = \lambda\mathbf{x}$ is fundamental.
- Classical algorithms such as GMRES (for linear systems) and Rayleigh–Ritz / eigs (for eigenvalue problems) are accurate but costly for large n.
- **Randomized algorithms** (e.g., sketching) enable fast dimension reduction, making traditional solvers scalable.

Goal: Develop projection-based solvers accelerated by random sketching, retaining accuracy while reducing cost.

Sketching: A Primer

- Sketching: project high-dimensional problem onto a low-dimensional subspace.
- Use a random matrix $\mathbf{S} \in \mathbb{C}^{s \times n}$ $(s \ll n)$ such that:

$$\mathbb{E}\|\mathbf{S}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \mathbb{C}^n$$

- Leads to small least-squares or eigenvalue subproblems.

Sketching + GMRES (sGMRES)

- Classic GMRES solves $\mathbf{Ax} = \mathbf{f}$ by finding an approximate solution $\mathbf{x}_B = \mathbf{By}$ in the Krylov subspace $\mathcal{K}_d(\mathbf{A}, \mathbf{f})$.
- It minimizes the residual $\|\mathbf{Ax}_B - \mathbf{f}\|_2$ by solving the least-squares problem:

$$\min_{\mathbf{y} \in \mathbb{C}^d} \|\mathbf{ABy} - \mathbf{f}\|_2$$

- sGMRES replaces this with a sketched version:

$$\min_{\mathbf{y} \in \mathbb{C}^d} \|\mathbf{S}(\mathbf{ABy} - \mathbf{f})\|_2$$

- $\mathbf{B}$: Krylov or truncated Arnoldi basis; $\mathbf{S}$: sketching matrix.

Krylov Basis Construction (Truncated Arnoldi)

- GMRES uses the Arnoldi process to build an orthonormal basis for the Krylov subspace:

$$\mathcal{K}_d(\mathbf{A}, \mathbf{f}) = \text{span}\{\mathbf{f}, \mathbf{A}\mathbf{f}, \mathbf{A}^2\mathbf{f}, \dots\}$$

- Full orthogonalization is expensive for large d.
- Truncated Arnoldi avoids full orthogonalization to save time:

$$\mathbf{b}_j = \frac{(I - \sum_{i=j-k}^{j-1} \mathbf{b}_i \mathbf{b}_i^*)\mathbf{A}\mathbf{b}_{j-1}}{\| \cdot \|_2}$$

- Works well in practice even with $k = 2$ or 4.

Theoretical Guarantees: sGMRES

- With high probability, the sketching matrix $\mathbf{S} \in \mathbb{C}^{s \times n}$ (e.g., Gaussian, SRHT) satisfies:

$$(1-\varepsilon)\|\mathbf{r}\|_2 \leq \|\mathbf{S}\mathbf{r}\|_2 \leq (1+\varepsilon)\|\mathbf{r}\|_2 \quad \forall \mathbf{r} \in \text{range}(\mathbf{AB})$$

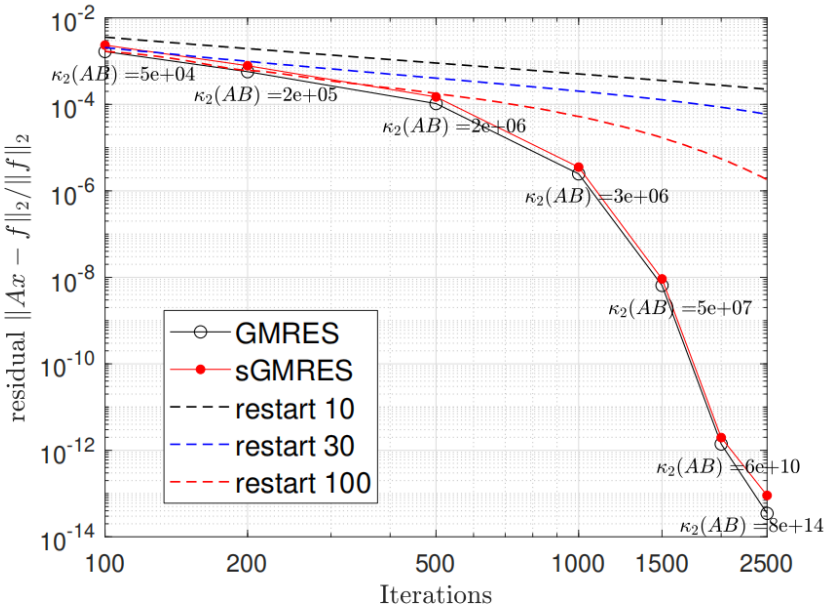
- Implies approximate solution $\mathbf{x}_B = \mathbf{B}\mathbf{y}^\star$ from sGMRES satisfies:

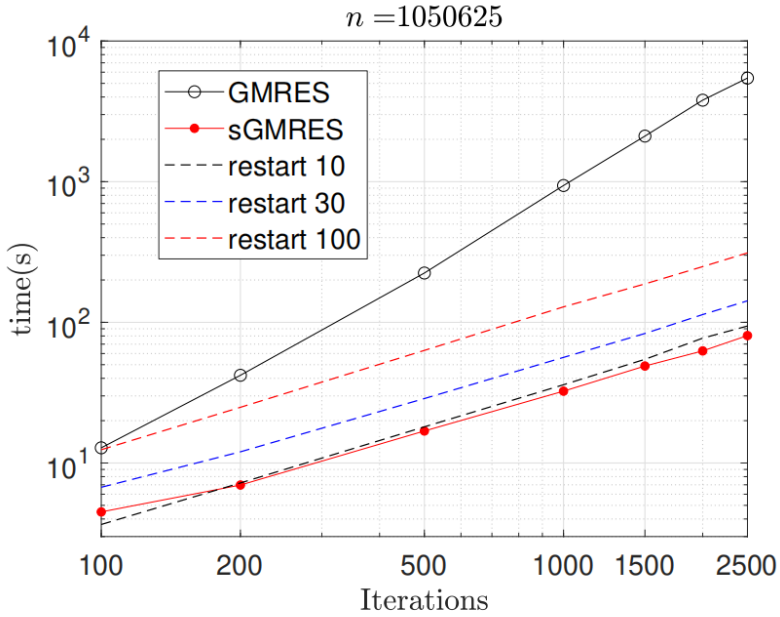
$$\|\mathbf{A}\mathbf{x}_B - \mathbf{f}\|_2 \leq (1+\varepsilon) \min_{\mathbf{y}} \|\mathbf{A}\mathbf{B}\mathbf{y} - \mathbf{f}\|_2$$

- Sketch size $s = \mathcal{O}(d \log d)$ suffices for $(1 \pm \varepsilon)$-accuracy with high probability.

Comparison: GMRES vs sGMRES

- **GMRES:** $\mathcal{O}(nd^2)$ operations
- **sGMRES:** $\mathcal{O}(d^3 + nd\log d)$ operations
- Up to 70× faster for PDE discretizations (e.g., convection-diffusion)





Sketching + Rayleigh–Ritz (sRR)

- Solve $\mathbf{Ax} = \lambda\mathbf{x}$ via sketched Rayleigh–Ritz:

$$\min_{\mathbf{M}\in\mathbb{C}^{d\times d}} \|\mathbf{S}(\mathbf{AB} - \mathbf{BM})\|_F$$

- Compute $\widehat{\mathbf{M}} = (\mathbf{SB})^{\dagger}(\mathbf{SAB})$ then eigendecompose.
- Faster and accurate even for poorly conditioned $\mathbf{B}$.

Theoretical Guarantees

- Suppose $\mathbf{B} \in \mathbb{C}^{n \times d}$ spans a good approximate invariant subspace of $\mathbf{A} \in \mathbb{C}^{n \times n}$.
- Let $\mathbf{S} \in \mathbb{C}^{s \times n}$ be a random matrix with i.i.d. sub-Gaussian entries (or SRHT), and define:

$$\widehat{\mathbf{M}} = (\mathbf{SB})^{\dagger}(\mathbf{SAB})$$

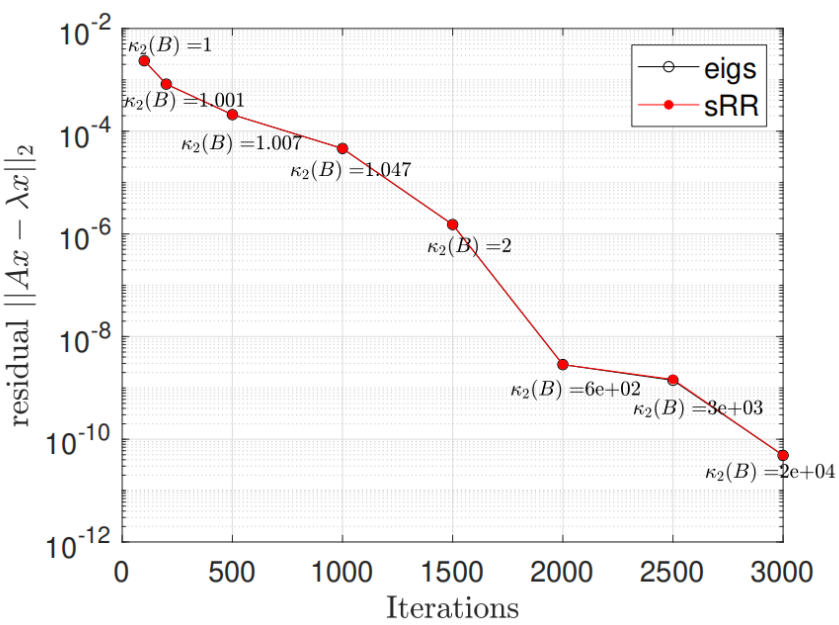
- Then for any $0 < \varepsilon < 1$, if $s \geq C \cdot d \log(d/\delta)/\varepsilon^2$, we have with probability at least $1 - \delta$:

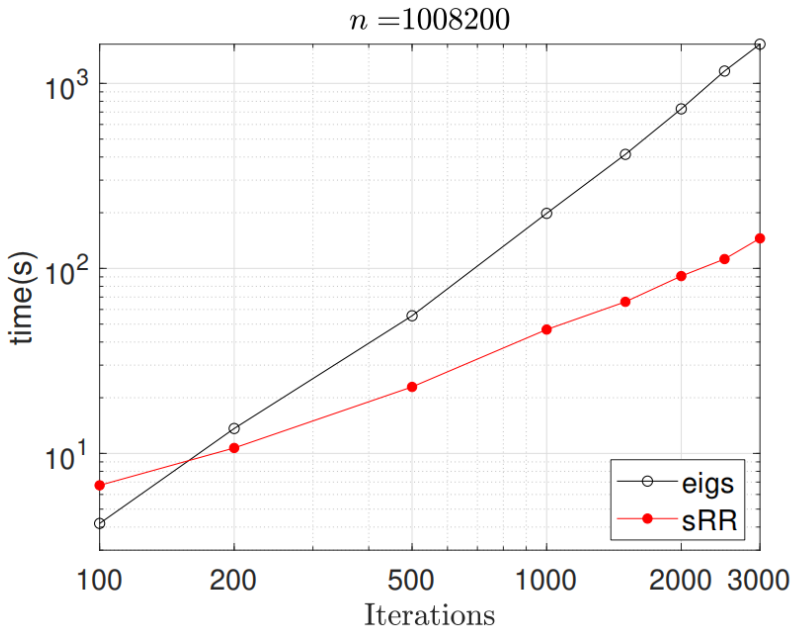
$$\left\| \widehat{\mathbf{M}} - \mathbf{B}^* \mathbf{AB} \right\| \leq \varepsilon \|\mathbf{A}\|_2$$

- Consequently, the eigenvalues of $\widehat{\mathbf{M}}$ approximate those of $\mathbf{A}$ in range$(\mathbf{B})$ up to $\mathcal{O}(\varepsilon)$ error.

Comparison: eigs vs sRR

- **eigs:** classic Arnoldi + RR, expensive orthogonalization
- **sRR:** fast basis + sketching = 10× speedup
- Accuracy preserved, suitable for optimization subproblems





Advantages and Limitations

Advantages

- Asymptotically faster than classic GMRES/RR
- Flexibility in basis choice
- Simple and practical implementation

Advantages and Limitations

Advantages

- Asymptotically faster than classic GMRES/RR
- Flexibility in basis choice
- Simple and practical implementation

Limitations

- Needs numerically full-rank basis
- Performance relies on good sketch and basis construction

Conclusion

- Introduced **sGMRES** and **sRR**: fast randomized solvers for linear systems and eigenvalue problems.
- Sketching enables dimension reduction with minimal loss of accuracy.
- Promising direction for large-scale scientific computing.