Science and Technology Facilities Council
Hartree Centre

Paper

hartree.stfc.ac.uk

Fast Linear Solvers via AI-Tuned Markov Chain Monte Carlo-based Matrix Inversion

Anton Lebedev[1*], Won Kyung Lee[1*], Soumyadip Ghosh[2], Olha I. Yaman[1], Vassilis Kalantzis[2], Yingdong Lu[2], Tomasz Nowicki[2], Shashanka Ubaru[2], Lior Horesh[2], Vassil Alexandrov[1]
[1] STFC Hartree Centre, Sci-Tech Daresbury, Warrington, UK; [2] IBM Research, Yorktown Heights, NY, USA; * Equal contribution

Introduction

- Large, sparse linear systems $Ax = b$ are pervasive in modern science/engineering. Krylov solvers converge slowly on ill-conditioned matrices.
- Preconditioning: Build a sparse, low-cost approximate inverse P, solve $PAx = Pb \rightarrow$ Fewer Krylov iterations.
- MCMC matrix inversion (MI): Generates P via many random walks; parallel but sensitive to a few MCMC algorithmic parameters. Grid search → high-cost
- Goal: Recommend MCMC parameters for a given matrix A, reducing iterations under a tight evaluation budget.
- Approach: Bayesian Optimisation (BO) with
  - A Graph Neural Network (GNN)-based surrogate that predicts a preconditioning metric $y(A, x_M)$ with uncertainty
  - Expected Improvement (EI) acquisition function selecting the next candidates
- Results: On an unseen, ill-conditioned system, BO achieves **~10% fewer Krylov iterations** with **50% of the search budget**.

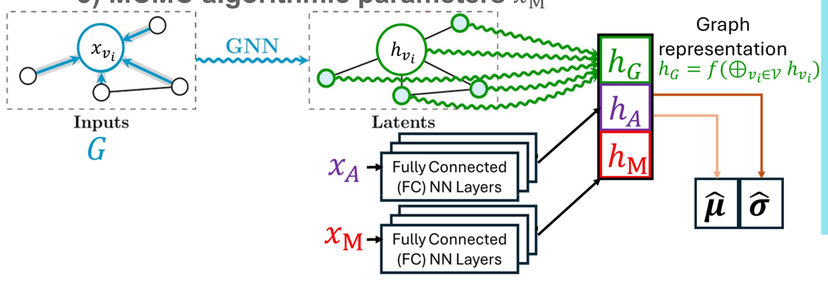
Method

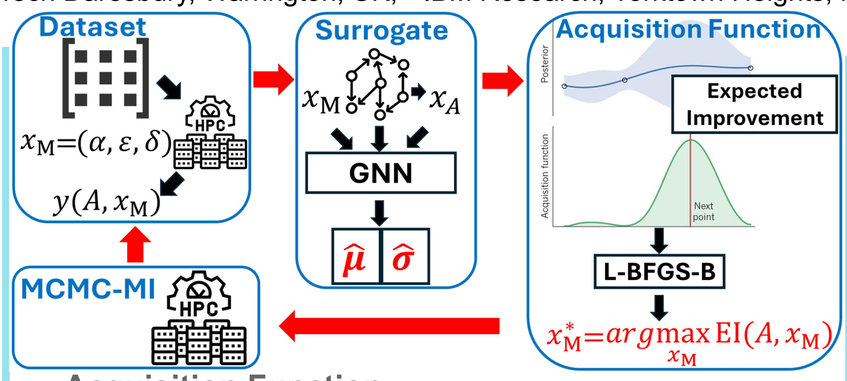
Graph Neural Surrogate

Inputs **1) Graph G from A**
- rows as nodes
- nonzeros A_{ij} as weighted edges
- node degree as node attribute

2) Inexpensive matrix features x_A
- 1-/Frobenius-/Infinity norms
- Largest/Smallest nonzeros; Sparsity, Trace, Symmetricity

3) MCMC algorithmic parameters x_M



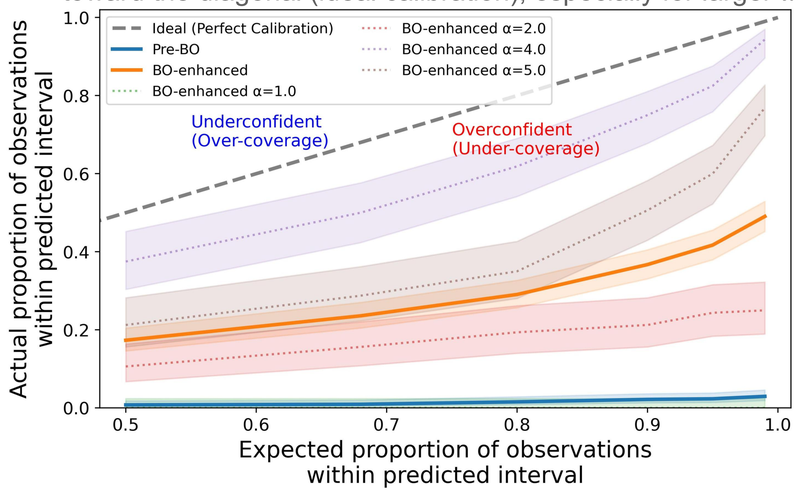


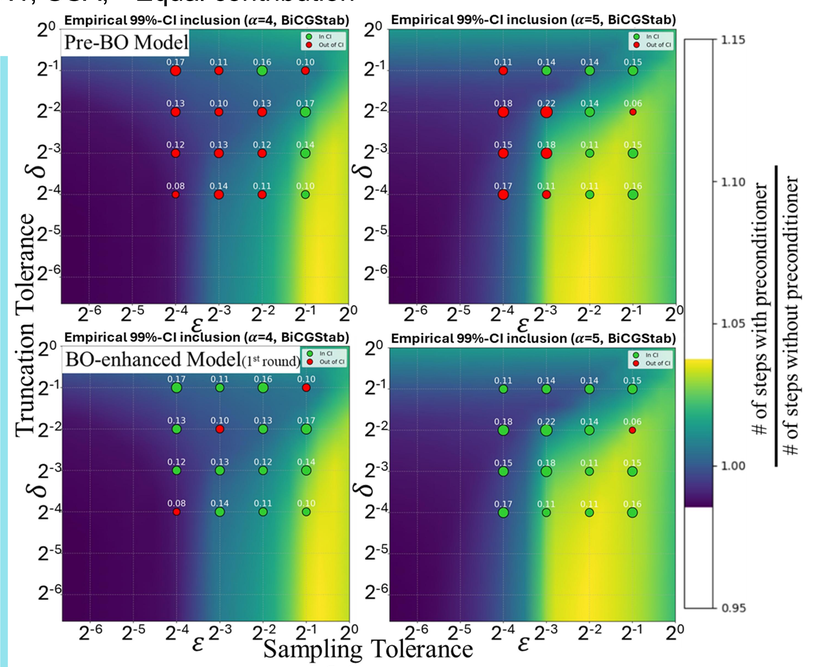
Acquisition Function

Expected Improvement
- to balance exploitation / exploration
- gradient-based optimisation (**L-BFGS-B, multi-start**) yields batched proposals

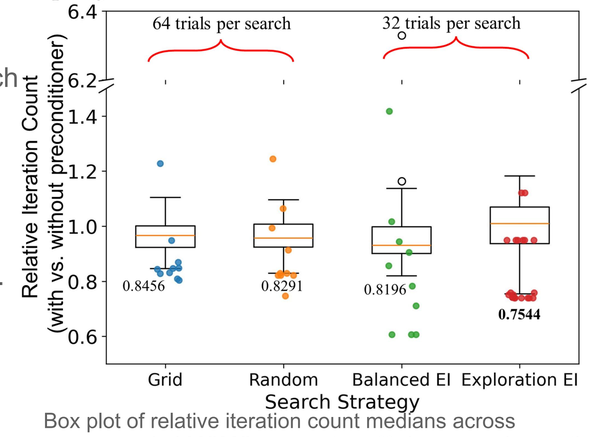
Experiments & Findings

- **MCMCMI parameters**, [1,2]
  $\alpha > 0$: diagonal shift ensuring Neumann–series convergence
  $\varepsilon \in (0,1]$: sampling tolerance (controls effective chain count)
  $\delta \in (0,1]$: truncation tolerance (caps maximum walk length)
- **Training Data**:
  11 PDE-derived sparse matrices from plasma physics to climate simulation (64-20,000 dim.)
  64-point grid over x_M with 10 replicates
  80/20 train/validation
- **Testing Data**
  ill-conditioned higher-order discretisation of advection–diffusion matrix ($\kappa \approx 6.6 \times 10^6$)
- After one EI round (32 evals) and retraining, curves move toward the diagonal (ideal calibration), especially for larger α.





- With 50% of search budget, BO finds better MCMCMI parameters x_M delivering **~10% fewer Krylov iterations**.



Box plot of relative iteration count medians across recommended MCMC parameters. shows an observation under the best parameter yielding the lowest median.

Outlook

- **Optimise time-to-solution in realistic HPC environments**
  GPU-accelerated and multi-node systems
  accounting for latency, communication & memory overheads
- **Strengthen surrogate** model with deep kernels or scalable GPs
- Upgrade the acquisition scheme (**cost-aware, batch, constrained**)
- **Active learning loop** with Generative AI (Exploring informative A)

Acknowledgements

This work was supported by the Hartree National Centre for Digital Innovation, a UK Government-funded collaboration between STFC and IBM.

References

[1] Lebedev et al. (2018). On advanced Monte Carlo methods for linear algebra on advanced accelerator architectures. In ScalA Workshop@SC18.
[2] Sahin et al. (2021). Usability of Markov chain Monte Carlo preconditioners in practical problems. In ScalA Workshop@SC21.