Randomized Numerical Linear Algebra
Poster Session Abstracts
February 3, 2026
4:00-5:00pm

A High Performance GPU CountSketch Implementation and Application to Least Squares Problems
Andrew Higgins, Sandia National Labs

Random sketching is a dimensionality reduction technique that approximately preserves norms and singular values up to some $O(1)$ distortion factor with high probability. The most popular sketches in literature are the Gaussian sketch and the subsampled randomized Hadamard transform, while the CountSketch has lower complexity. Combining two sketches, known as multisketching, offers an inexpensive means of quickly reducing the dimension of a matrix by combining a CountSketch and Gaussian sketch. However, there has been little investigation into high performance CountSketch implementations. In this work, we develop an efficient GPU implementation of the CountSketch, and demonstrate the performance of multisketching using this technique. We also demonstrate the potential for using this implementation within a multisketched least squares solver that is up to 77% faster than the normal equations with significantly better numerical stability, at the cost of an $O(1)$ multiplicative factor introduced into the relative residual norm.

Recursive Sketched Interpolation: Efficient Hadamard Products of Tensor Trains
Zhaonan Meng, North Carolina State University

The Hadamard (element-wise) product of two tensors represented in the tensor-train (TT) format is an important operation in many TT-based applications, including the solution of nonlinear differential equations and fast convolution of tensorized functions. However, conventional approaches to computing this product often scale poorly with the bond dimensions (TT-ranks), creating a computational bottleneck in practice. By combining tensor-train sketching with slice selection via interpolative decomposition, we introduce a “scale product” algorithm, Recursive Sketched Interpolation (RSI). The algorithm computes the Hadamard product of TTs with bond dimension χ at a computational cost of O(χ^3). By evaluating RSI against conventional baseline methods for TTs in diverse scenarios, we demonstrate its superior scalability while maintaining comparable accuracy.

Subspace-constrained randomized coordinate descent for linear systems with good low-rank matrix approximations
Jackie Lok, Princeton University

We describe an efficient linear solver for solving positive semidefinite linear systems with approximate low-rank structure. The algorithm is based on restricting the dynamics of the classical randomized block coordinate descent method within an affine subspace corresponding to a column Nyström approximation, efficiently computed using an algorithm for low-rank matrix approximation such as the recently proposed RPCholesky algorithm. The convergence rate of this method is unaffected by large spectral outliers, making it an effective and memory-efficient solver for large-scale, dense linear systems with rapidly decaying spectra, such as those encountered in kernel ridge regression. The theoretical results are derived by developing a more general subspace-constrained framework for the sketch-and-project method, which provides a flexible, implicit preconditioning strategy for a wider range of iterative solvers.

Boosting VarProNets: Efficient Gradient Boosting with Variable Projection
Abhijit Chowdhary, Tufts University

Deep neural networks (DNNs) are a versatile framework for a wide range of machine learning tasks, from image classification to approximating physical processes. However, training these models remains a critical challenge, particularly in high-dimensional problems where it can be unreliable and computationally expensive. We seek to develop a more robust and efficient training paradigm by combining two modern techniques: Variable Projection (VarPro) and gradient boosting. Our approach improves gradient boosting, a greedy ensemble training technique, by exploiting the linear structure within each learner using VarPro. By integrating these methods, we aim to create a provably more reliable and less expensive training framework for neural networks. This work is accompanied with benchmarks against the state-of-the-art in gradient boosting.

Efficient QR-based Column Subset Selection through Randomized Sparse Embeddings
Israa Fakih, PSI/EPFL

The poster introduces an efficient algorithm for column subset selection that combines the column-pivoted QR factorization with sparse subspace embeddings. The proposed method, SE-QRCS, is particularly effective for wide matrices with significantly more columns than rows. Starting from a matrix A, the algorithm selects k columns from the sketched matrix B = A *S^T, where S is a sparse subspace embedding of range(A^T). The sparsity structure of S is then exploited to map the selected pivots back to the corresponding columns of A, which are then used to produce the final subset of selected columns. We prove that this procedure yields a factorization with strong rank-revealing properties. The resulting bounds exhibit a reduced dependence on the number of columns of A compared to those obtained from the traditional strong rank-revealing QR factorization. Moreover, when the leverage scores are known the bounds become entirely independent of the column dimension. For general matrices, the algorithm can be extended by first applying an additional subspace embedding of range(A).

Iterative low-rank time integration of the Schrödinger equation
Polina Sachsenmaier, RWTH Aachen University

Standard numerical methods for solving PDEs typically suffer from the curse of dimensionality: their computational cost scales exponentially with the dimension of the underlying domain, making them impractical even at low resolution. In many cases of interest, however, such limitations can be overcome by appropriate compressed representations of approximate solutions, in particular by low-rank tensor representations.For time-dependent PDEs, several approaches to low-rank approximation exist that control ranks in different ways. These range from methods that keep ranks fixed, such as dynamical low-rank approximations, which may lead to uncontrolled errors, to methods that approximate standard time-stepping schemes to any desired accuracy but can produce unnecessarily large ranks. We develop time-integration methods in low-rank tensor representations that adaptively adjust approximation ranks to meet a prescribed accuracy while simultaneously ensuring control over the ranks of the computed approximations and all intermediate quantities, which are also a main determining factor in the computational costs of such methods. Our approach combines an iterative time-stepping scheme with soft thresholding of the iterates. In the matrix case, the proposed strategy yields iterates whose ranks remain comparable to natural benchmark quantities — namely, the best approximation ranks of the sought solutions at the achieved accuracy. In the higher-dimensional tensor case the algorithm can be adapted appropriately, leading to global error and rank bounds that depend only polynomially on the dimension. Numerical experiments illustrate the theory for linear time-dependent Schrödinger equations. This poster is based on joint works with Markus Bachmayr, Matthieu Dolbeault, Tianyu Jin and Federico Vismara.

A Practical Mode-Parallel Implementation Of The (H-)Tucker Decomposition Via Randomization
Sascha Portaro, University of Bologna

Tensors provide a natural and compact representation for multidimensional data, and low-rank factorizations are often key to unveiling hidden structures arising from dependencies among variables. However, computing such factorizations becomes increasingly demanding in terms of memory and energy consumption as the tensor order grows. We focus on two state-of-the-art tensor decompositions, namely Tucker and H-Tucker, and introduce novel numerical strategies that enable a fully mode-parallel computation, where operations along all tensor modes are performed simultaneously rather than sequentially. Our approach leverages modern randomization techniques, including fiber sampling and randomized range-finding, to improve computational efficiency. We establish upper bounds on the expected approximation error and demonstrate through numerical experiments significant reductions in runtime and memory usage. Additional experiments in high-performance computing environments highlight the good scalability of the proposed mode-parallel framework.

Parametric Hierarchical Matrix Approximations to Kernel Matrices
Abraham Khan, North Carolina State University

Kernel matrices arising in applications such as Gaussian processes may not always admit a low-rank approximation. Important examples are kernel matrices induced by certain members of the Matérn family of covariance kernels, with smaller length scales and values of ν. Still, they can often be approximated by a hierarchical matrix ($\mathcal{H}$-matrix or $\mathcal{H}^{2}$-matrix), which consists of a hierarchy of small near-field blocks (sub-matrices) stored in a dense format and large low-rank far-field blocks that are efficiently stored in factored form. A hierarchical matrix approximation of a kernel matrix can be constructed, stored, and used to perform matrix-vector multiplication in log-linear or linear complexity with respect to n. Standard methods for approximating kernel matrices with hierarchical matrices do not account for the following: kernel matrices often depend on certain hyperparameters that must be optimized over a fixed parameter space. For example, in Gaussian processes and Bayesian inverse problems, estimating the hyperparameters from the data involves solving an optimization problem, which requires repeatedly forming or approximating the kernel matrices for a range of parameters. To address this computational challenge, we introduce a new class of hierarchical matrices, namely, parametric (parameter-dependent) hierarchical matrices. The construction of a parametric hierarchical matrix follows an offline-online paradigm. In the offline stage, the near-field and far-field blocks are approximated by using polynomial approximation and tensor compression. In the fast online stage, for a particular hyperparameter, the parametric hierarchical matrix is instantiated efficiently as a standard hierarchical matrix. Numerical experiments show speedups of over 100 times compared with existing techniques.

Structured Pseudospectral Divide-and-Conquer for Definite Pencils
Ryan Schneider, University of California Berkeley

We demonstrate how to adapt the fastest algorithm for the generalized eigenvalue problem (pseudospectral divide-and-conquer) to definite pencils. Our new implementation – which preserves definiteness – is both easier to use and asymptotically faster on definite problems. The key insight underpinning the algorithm is a new symmetrized version of pseudospectral shattering, which describes the regularizing effect of structured perturbations (either random diagonal or sampled from the Gaussian unitary ensemble) on the spectrum and pseudospectrum of a definite pencil. Stop by this poster to see what symmetrized pseudospectral shattering looks like, find out what preserving structure buys us in spectral divide-and-conquer, and learn why randomized eigensolvers (like this one) deserve to be added to our standard software libraries. Based on joint work with James Demmel and Ioana Dumitriu.

Adaptive LSQR Preconditioning from One Small Sketch
Jung Eun, University of Oxford

Large-scale linear least squares problems arise in many areas of computational science and data analysis, where efficiency and scalability are crucial. In this talk, we introduce a randomized preconditioning framework for iterative solvers based on low-rank approximations of small sketches of the original problem. The key idea is to iteratively construct low-rank preconditioners that reshape the singular value distribution in a favorable way. By tightly coupling the preconditioning and Krylov solving phases within an iterative CUR decomposition -- a low-rank approximation built from selected of columns and rows of the original matrix -- the proposed algorithm achieves faster and earlier convergence than existing methods. The algorithm performs particularly well on problems that are large in both dimensions, as well as on sparse and ill-conditioned systems. This is a joint work with Coralia Cartis and Yuji Nakatsukasa.

Faster Linear Algebra Algorithms with Structured Random Matrices
Chris Camaño, Caltech

To achieve the greatest possible speed, practitioners regularly implement randomized algorithms for low-rank approximation and least-squares regression with structured dimension reduction maps. Despite significant research effort, basic questions remain about the design and analysis of randomized linear algebra algorithms that employ structured random matrices. This poster presents a new theoretical framework for studying structured random matrices, and details various algorithm acceleration techniques based on sketching with random matrices possessing sparse or tensor product structure.

Matrix analysis for shallow ReLU neural network least-squares approximations
Tong Ding, Purdue University

Neural network provides an effective tool for the approximation of some challenging functions. However, fast and accurate solvers for relevant dense linear systems are rarely studied. This work gives a comprehensive characterization of the ill conditioning of some dense linear systems arising from shallow neural network least squares approximations. It shows that the systems are typically very ill conditioned, and the conditioning gets even worse with challenging functions such as those with jumps. This makes the solutions hard for typical iterative solvers. On the other hand, we can further show the existence of some intrinsic rank structures within those matrices, which make it feasible to obtain nearly linear complexity robust direct solutions. Most of our discussions focus on the 1D case, but extensions to some 2D cases are also given.

Adaptive matrix approximations for PDE-Constrained Inverse Problems
Harshit Bhatt, North Carolina State University

Large-scale inverse problems governed by PDEs require repeated access to Hessian (or Gauss-Newton) matrices. Assembling and storing these are infeasible in large-scale settings and can be as costly as the cubic order of the dimensions when possible. Our goal is to exploit the hierarchical off-diagonal low-rank (HODLR) structure of these matrices and to formulate adaptive randomized schemes that achieve log-linear storage and fast factorizations, thereby attaining the desired approximation accuracy without any prior knowledge.

A Convergent Generalized Krylov Subspace Method for Compressed Sensing MRI Reconstruction with Gradient-Driven Denoisers
Tao Hong, University of Texas at Austin

Model-based reconstruction plays a key role in compressed sensing (CS) MRI, as it incorporates effective image regularizers to improve the quality of reconstruction. The Plug-and-Play and Regularization-by-Denoising frameworks leverage advanced denoisers (e.g., convolutional neural network (CNN)-based denoisers) and have demonstrated strong empirical performance. However, their theoretical guarantees remain limited, as practical CNNs often violate key assumptions. In contrast, gradient-driven denoisers achieve competitive performance, and the required assumptions for theoretical analysis are easily satisfied. However, solving the associated optimization problem remains computationally demanding. To address this challenge, we propose a generalized Krylov subspace method (GKSM) to solve the optimization problem efficiently. Moreover, we also establish rigorous convergence guarantees for GKSM in nonconvex settings. Numerical experiments on CS MRI reconstruction with spiral and radial acquisitions validate both the computational efficiency of GKSM and the accuracy of the theoretical predictions. The proposed optimization method is applicable to any linear inverse problem.

Quantile Randomized Kaczmarz with Time-Varying Noise in Signal
Emeric Battaglia, University of California, Irvine

When solving a linear system of equations, the signal vector may be corrupted by time-varying noise. When the corruption is sparse, quantile-based methods such as the quantile Randomized Kaczmarz method (qRK) can recover the underlying solution to the system. However, when the corruption is insufficiently sparse, convergence is only guaranteed up to an error horizon. In the setting of time-varying noise, current theory also suggests convergence up to an error horizon. However, if the noise decays over time, one would hope that qRK can recover the solution to the system. In this work, we present an improved error horizon bound and demonstrate that it improves the previously known bounds under mild constraints. We also demonstrate that, in the time-varying case, the horizon may decay as well, and qRK converges to the unique solution of the system.

Stochastic gradient with least-squares control variates
Matteo Raviola, EPFL

The stochastic gradient (SG) method is a widely used approach for solving stochastic optimization problems, but its convergence is typically slow. Existing variance reduction techniques, such as SAGA, improve convergence by leveraging stored gradient information; however, they are restricted to settings where the objective functional is a finite sum, and their performance degrades when the number of terms in the sum is large. In this work, we propose a novel approach which also works when the objective is given by an expectation over random variables with a continuous probability distribution. Our method constructs a control variate by fitting a linear model to past gradient evaluations using weighted discrete least-squares, effectively reducing variance while preserving computational efficiency. We establish theoretical sublinear convergence guarantees and demonstrate the method's effectiveness through numerical experiments on random PDE-constrained optimization.

Adjoint-free operator learning and transpose-free linear algebra
Diana Halikias, Courant Institute, New York University

There is a mystery at the heart of operator learning: how can one recover a non-self-adjoint operator from data without probing the adjoint? Current practical approaches suggest that one can accurately recover an operator while

only using data generated by the forward action of the operator without access to the adjoint. However, naively, it seems essential to sample the action of the adjoint. We partially explain this mystery by proving that without querying the adjoint, one can approximate a family of non-self-adjoint infinite-dimensional compact operators via projection onto a Fourier basis. We then apply the result to recovering Green's functions of elliptic partial differential operators and derive an adjoint-free sample complexity bound. While existing theory justifies low sample complexity in operator learning, ours is the first adjoint-free analysis that attempts to close the gap between theory and practice. We also provide an accompanying analysis of transpose-free low-rank approximation.

Quasi-optimal hierarchically semi-separable matrix approximation
David Persson, New York University & Flatiron Institute

We present a randomized algorithm for producing a quasi-optimal hierarchically semi-separable (HSS) approximation to an $N\times N$ matrix A using only matrix-vector products with A and A^T. We prove that, using $O(k \log(N/k))$ matrix-vector products and $O(N k^2 \log(N/k))$ additional runtime, the algorithm returns an HSS matrix B with rank-k blocks whose expected Frobenius norm error $\mathbb{E}[\|A - B\|_F^2]$ is at most $O(\log(N/k))$ times worse than the best possible approximation error by an HSS rank-k matrix. In fact, the algorithm we analyze in a simple modification of an empirically effective method proposed by [Levitt & Martinsson, SISC 2024]. As a stepping stone towards our main result, we prove two results that are of independent interest: a similar guarantee for a variant of the algorithm which accesses A's entries directly, and explicit error bounds for near-optimal subspace approximation using projection-cost-preserving sketches. To the best of our knowledge, our analysis constitutes the first polynomial-time quasi-optimality result for HSS matrix approximation, both in the explicit access model and the matrix-vector product query model.