Abstracts from Workshop on Sparse Tensor Computations

18-19 October 2023

Day 1 18-October-2023, Session 1, Time: 9:00-10:30am EST

Title Low rank Tensor Approximations to Kinetic Models
Speaker: Jingmei Qiu, University of Delaware, jingqiu@udel.edu
Abstract: We propose a novel Local Macroscopic Conservative (LoMaC) low rank tensor method with discontinuous Galerkin (DG) discretization for the physical and phase spaces for simulating the Vlasov-Poisson (VP) system. The LoMaC property refers to the exact local conservation of macroscopic mass, momentum and energy at the discrete level. The LoMaC low rank tensor algorithm (recently developed in arXiv:2207.00518) simultaneously evolves the macroscopic conservation laws of mass, momentum and energy using the kinetic flux vector splitting; then the LoMaC property is realized by projecting the low rank kinetic solution onto a subspace that shares the same macroscopic observables. This paper is a generalization of our previous work, but with DG discretization to take advantage of its compactness and flexibility in handling boundary conditions and its superior accuracy in the long term. Extensive numerical results are performed to showcase the efficacy of the method

Title Tensor Computations in Lattice QCD
Speaker: William Detmold, MIT, wdetmold@mit.edu
Abstract: Lattice Quantum Chromodynamics calculations are central many areas of particle and nuclear physics, begin necessary for many comparisons between experiment and the theoretical framework of the Standard Model. In this talk, I will briefly introduce the technique of LQCD and the way in which tensor contractions enter into these calculations for complicated systems such as nuclei. I will highlight a few approaches that are used within the field and some possibilities for future directions.

Title Federated tensor learning and its application for healthcare
Speaker: Joyce Ho, Emory University, joyce.c.ho@emory.edu
Abstract: Tensor factorization can provide a data-driven framework for learning representations in health data analysis, especially for computational phenotyping where the high-dimensional electronic health records can be converted to meaningful and interpretable medical concepts. Unfortunately, the sensitive nature of the health data restricts collaboratively learning phenotypes from different institutes (e.g., hospitals, clinics, pharmaceutical companies, and health insurance providers). Federated tensor factorization, a special distributed tensor factorization paradigm, enables parallelization of the tensor computation but also preserves data privacy by avoiding direct data sharing, and learns the shared phenotypes through joint tensor factorization without communicating the individual-level data. We propose a communication-efficient federated generalized tensor factorization framework that can adapt to different types of data and reduces the communication cost without compromising convergence and factorization quality. We demonstrate our framework in two different federated learning scenarios: a central parameter server and the decentralized setting.

Title Automating symbolic manipulation and evaluation of data-sparse tensor algebra for quantum electronic structure

Speaker: Ed Valeev, Virginia Tech, efv@vt.edu

Abstract: Even approximate quantum mechanics of indistinguishable particles gives rise naturally to the tensor algebra of symmetric/antisymmetric tensors of high order (≥ 6). For electrons in ordinary matter these tensors have exploitable data-sparse structure that allows to formulate predictive quantum computational methods that approach the classical, N log N, complexity for the N-body system. The gamut of the data-sparsity in these tensors runs from ordinary element and block sparsity to the more complex block-rank-sparse patterns. Whereas even the former pose significant formal and computational challenges (e.g., error control, scalable algorithms, etc.), with toolchains for such compute patterns only recently emerging, the block-rank-sparse tensor structures are an essential new frontier for tensor computation. To simplify the implementation of reduced-complexity many-body electronic structure formalisms featuring data-sparse tensors (including the block-rank-sparse tensors) we have been developing support for symbolic manipulation of tensor networks involving data-compressed tensors as well as their automated evaluation, using two open-source frameworks developed by our group, namely the SeQuant computer algebra system (GitHub.com:ValeevGroup/SeQuant) and the TiledArray tensor framework ((GitHub.com:ValeevGroup/TiledArray). In this talk I will briefly review some of these developments

Day 1 18-October-2023, Session 2, Time: 11:00-Noon EST

Title: Compiler Support for Structured Data

Speaker: Changwan Hong, Massachusetts Institute of Technology (MIT), changwan@csail.mit.edu

Abstract: In 1957, FORTRAN revolutionized programming by introducing multi-dimensional dense arrays or dense tensors. Over the years, programming languages expanded to include a diverse array of data structures like lists, sets, hash tables, trees, and graphs. Despite this diversity, dense tensors remained a vital and practical solution for handling large datasets. However, modern data is often sparse and structured. Real-world data, whether originating from sensors, computations, or human activity, often exhibits underlying structures such as sparsity, repeated patterns, or symmetry.

In this talk, I will present our approaches to address the challenges posed by structured and sparse modern datasets. First, I will present innovative abstractions aimed at representing losslessly compressed data as a generalized form of sparse tensors, enriched with repetitions of values. Second, I will introduce the Looplet language, designed to efficiently abstract array structures, enabling the compiler to seamlessly generate code by coiterating over any combination of diverse array structures. Third, I will talk about our automated asymptotic scheduler designed specifically for sparse tensor programs. This scheduler allows us to represent the asymptotic cost of schedules, assisting in making informed decisions when choosing between different schedules. Lastly, I will present WACO, a tool which simultaneously optimizes the format and processing schedules of sparse tensor programs. WACO leverages a sparse convolutional network to comprehend the significant features of sparsity patterns.

Title: Linearized Tensor Format for Performance-Portable Sparse Tensor Computation

Speaker: Jee Choi, University of Oregon, jeec@uoregon.edu

Abstract: We demonstrate the efficiency and performance-portability of encoding sparse tensors in a linearized format. This encoding scheme allows a unified algorithm for matricized tensor times Khatri-Rao product (MTTKRP) that achieves high performance on both CPUs and GPUs, using a single, compact, and mode-agnostic copy of the tensor. Our new linearized sparse tensor format achieves geometric mean speedup of 4.3x and 2.6x over the state-of-the-art on the latest CPU and GPU architectures, respectively.

Title: Extending Einsums to Support Graph Analytics: A BFS Example

Speaker: Toluwanimi Odemuyiwa, University of California - Davis, todemuyiwa@ucdavis.edu

Abstract: Graph analytics, which seeks to understand the relationship between entities in a graph, is a broad, but important domain. For graph implementations, there is generally no separation of concerns between the computation being performed, and the software features developers leverage to enable that computation.

The diversity in algorithm categories and implementations poses a few challenges: (1) It is difficult to reason about computation patterns that are common across various algorithms. (2) Exploring new algorithms or the mapping space for a particular problem requires luck and/or ingenuity. (3) Exploring different high-level optimizations for a particular implementation tends to require significant engineering effort.

In addressing these challenges, we take inspiration from Halide (Ragan-Kelley et al., 2013). In Halide, the algorithm (what is being computed) is cleanly separated from the schedule/mapping (how computation occurs). The tensor algebra world has pursued this approach as well: a common abstraction is Einstein summation notation (Einsums), which concisely expresses the iteration space over which computation takes place, but structuring the computation itself is not specified by the Einsum.

We believe that Einsum notation is the right abstraction for graph algorithms. Extending Einsum notation to support graph algorithms enables: (1) A declarative, algorithmic language that is decoupled from imperative mapping specifications for a given problem. This, in turn, allows us to leverage prior tensor algebra tools such as TACO (Kjolstad et al., 2017), Sparseloop (Wu et al., 2022), and TeAAL (Nayak et al., 2023) to explore the implementation space. (2) Design explorations—with the potential for automation—of different implementation choices. These explorations are based on algebraically manipulating the Einsum expressions and exploring the mapping options presented by the iteration space of the Einsum. (3) Separation of platform-dependent and platform-independent implementation choices, such that a single platform-independent layer can target multiple backends that map to different systems (including accelerators, CPUs, and GPUs).

Day 1 18-October-2023, Session 3, Time: 1:30-3:00pm EST

Title: A semi-implicit, low-rank DG method for a kinetic model of radiation emission and absorption

Speaker: Cory Hauck, Oak Ridge National Laboratory, hauck@ornl.gov

Abstract: In this presentation, I will summarize recent work on low-rank methods for a kinetic model of neutrino-matter interactions. The model itself is a simple linear ODE, but it highlights a key, generic feature of collisional kinetic equations, which is the need to simulate the dynamics over long time scales. In particular, recovery of the equilibrium limit is a critical requirement of any numerical method.

We investigate a discontinuous Galerkin (DG), dynamical low-rank approximation (DLRA) combined with an operator splitting time integration strategy that first updates the low-rank bases and then the coefficients of the numerical solution with respect to these new bases. Each step is performed implicitly so that large time steps can be used. We provide conditions on the method to ensure that the solution converges to the correct equilibrium, and we demonstrate the theory with a few numerical results.

Title: Optimal Matrix-Mimetic Tensor Algebras via Variable Projection
Speaker: Elizabeth Newman, Emory University, elizabeth.newman@emory.edu
Abstract: Many data are naturally represented as multiway arrays or tensors, and as a result, multilinear data analysis tools have revolutionized feature extraction and data compression. Despite the success of tensor-based approaches, fundamental linear algebra properties often break down in higher dimensions. Recent advances in matrix-mimetic tensor algebra in have made it possible to preserve linear algebraic properties and, as a result, to obtain optimal representations of multiway data. Matrix-mimeticity arises from interpreting tensors as t-linear operators, which in turn are parameterized by invertible linear transformations. The choice of transformation is critical to representation quality, and thus far, has been made heuristically. In this talk, we will learn data-dependent, orthogonal transformations by leveraging the optimality of matrix-mimetic representations. In particular, we will exploit the coupling between transformations and optimal tensor representations using variable projection. We will highlight the efficacy of our proposed approach on image compression and reduced order modeling tasks.

Title: HOQRI: Higher-order QR Iteration for Scalable Tucker Decomposition
Speaker: Kejun Huang, University of Florida, kejun.huang@ufl.edu
Abstract: We propose a new algorithm called higher-order QR iteration (HOQRI) for computing the Tucker decomposition of large and sparse tensors. Compared to the celebrated higher-order orthogonal iterations (HOOI), HOQRI relies on a simple orthogonalization step in each iteration rather than a more sophisticated singular value decomposition step as in HOOI. More importantly, when dealing with extremely large and sparse data tensors, HOQRI completely eliminates the intermediate memory explosion by defining a new sparse tensor operation called TTMcTC. Furthermore, recognizing that the orthonormal constraints form a Cartesian product of Stiefel manifolds, we introduce the framework of manifold optimization and show that HOQRI guarantees convergence to the set of stationary points. Numerical experiments on synthetic and real data showcase the effectiveness of HOQRI. As a side contribution, we apply the newly introduced manifold optimization framework to the classical HOOI algorithm and show that it also converges to the set of stationary points.

Title: Efficient Fine-tuning of pretrained machine learning models using Tensor Training
Speaker: Huan He, University of Pennsylvania, Huan.He@Pennmedicine.upenn.edu
Abstract: In this work, we leverage tensor train decomposition for fine-tuning pretrained machine learning models. We develop a parameter efficient fine-tuning framework that freezes the pre-trained model weights and injects trainable tensor decomposition matrices into each layer of the architecture, greatly reducing the number of trainable parameters for downstream tasks. This approach allows for adaptation to new domains or datasets even under constrained computational and memory resources. We demonstrate the superiority in two different scenarios: MLP and transformer based networks Our results are promising; not only does this method significantly lower the computational burden, but it also maintains competitive performance metrics when compared to more resource-intensive fine-tuning techniques.

Day 1 18-October-2023, Session 4, Time: 3:15-4:40pm EST

Title: TeAAL: A Declarative Framework for Modeling Sparse Tensor Accelerators
Speaker: Nandeeka Nayak, University of Illinois Urbana-Champaign, ndnayak2@illinois.edu

Abstract: Over the past few years, the explosion in sparse tensor algebra workloads has led to a corresponding rise in domain-specific accelerators to service them. Due to the irregularity present in sparse tensors, these accelerators employ a wide variety of novel solutions to achieve good performance. At the same time, prior work on design-flexible sparse accelerator modeling does not express this full range of design features, making it difficult to understand the impact of each design choice and compare or extend the state-of-the-art. To address this, we propose TeAAL: a language and simulator generator for the concise and precise specification and evaluation of sparse tensor algebra architectures. We use TeAAL to represent and evaluate four disparate state-of-the-art accelerators—ExTensor, Gamma, OuterSPACE, and SIGMA—and verify that it reproduces their performance with high accuracy. Finally, we demonstrate the potential of TeAAL as a tool for designing new accelerators by showing how it can be used to speed up vertex-centric programming accelerators—achieving 1.9× on BFS and 1.1× on SSSP over GraphDynS.

Title: Can tensor factorization help us shrink language models?
Speaker: Saday Sadayappan, University of Utah, saday@cs.utah.edu

Title: Indexed Streams: A Formal Intermediate Representation for Fused Contraction Programs
Speaker: Scott Kovach, Stanford University, dskovach@stanford.edu
Abstract: We introduce indexed streams, a formal operational model and intermediate representation that describes the fused execution of a contraction language that encompasses both sparse tensor algebra and relational algebra. We prove that the indexed stream model is correct with respect to a functional semantics. We also develop a compiler for contraction expressions that uses indexed streams as an intermediate representation. The compiler is only 540 lines of code, but we show that its performance can match both the TACO compiler for sparse tensor algebra and the SQLite and DuckDB query processing libraries for relational algebra.

Title: Tensor methods for parametric low-rank kernel approximations
Speaker: Arvind Saibaba, North Carolina State University, asaibab@ncsu.edu
Abstract: Kernels are ubiquitous in many areas of scientific computing such as Gaussian processes, integral equations, n-body problems. At the heart of a kernel method, is forming and computing with a kernel matrix corresponding to pairwise interactions between two sets of points. We consider the case where the kernel matrix is parameterized by a set of parameters, e.g., hyperparameters, and consider low-rank kernel approximations that can be easily computed over a range of parameters. We provide low-rank matrix approximations, by viewing the problem through the lens of tensor decompositions and present efficient methods using Tucker and Tensor Train. This is joint work with Rachel Minster, Abraham Khan (NC State), and Misha Kilmer (Tufts).

Day 2 19-October-2023, Session 1, Time: 8:30-10:00am EST

Title: Recent advances in sampling-based methods for tensor decomposition
Speaker: Osman Malik, Lawrence Berkeley National Laboratory, oamalik@lbl.gov
Abstract: Due to their multidimensional nature, tensors are inherently plagued by the curse of dimensionality. Indeed, simply storing an N-way tensor with each dimension equal to I requires I^N numbers. This exponential dependence on N makes its way into the cost of algorithms for decomposing tensors. In this talk, I will discuss sampling-based alternating least squares (ALS) methods for tensor decomposition. In particular, I will focus on recent efforts to tackle the exponential dependence on N, as well as very recent work that also improves the dependence on I. Ultimately, the goal of these efforts is to increase the scalability of tensor decomposition algorithms.

Title: Sampling Methods for the Canonical Polyadic Decomposition
Speaker: Carmeliza Navasca, University of Alabama at Birmingham, cnavasca@uab.edu
Abstract: The Alternating Least-Squares (ALS) is one of the most well known method for approximating factor matrices in the canonical polyadic decomposition. ALS is fast, but it has some drawbacks. In this talk, we address one of them. We describe some sampling methods (non-adaptive and adaptive) for ALS to improve speed, accuracy and memory space for large tensors. Numerical results will be provided.

Title: Efficient tensor network contraction algorithms
Speaker: Linjian Ma, University of Illinois at Urbana Champaign, lma16@illinois.edu
Abstract: A tensor network uses a set of small tensors to implicitly represent the structure of data that is expensive to form explicitly. In this talk, we introduce techniques to efficiently approximate an arbitrary tensor network into a single tensor. These techniques approximate large intermediate tensors as low-rank tree tensor networks, which reduces the memory usage and computational overhead for subsequent contractions. Our techniques involve algorithms that facilitate the selection of efficient contraction trees, the determination of tree tensor network structures and permutations, as well as the implementation of low-rank tree tensor network approximation algorithms. Experimental results indicate that the proposed technique outperforms previously proposed tensor network contraction algorithms for multiple problems in statistical physics and quantum computing in terms of both accuracy and efficiency.

Title: Faster Implicit Leverage Sampling Algorithms for CP and Tensor-Train Decomposition
Speaker: Vivek Bharadwaj, UC Berkeley, vivek_bharadwaj@berkeley.edu
Abstract: Randomized sketching algorithms have recently led to significant asymptotic and practical advances in tensor computations. Methods based on *statistical leverage scores* have contributed to many of these advances; multiple lines of work use leverage-based algorithms to achieve strong guarantees for Candecomp / PARAFAC (CP), Tucker, and Tensor-Train decomposition. In this talk, we describe improved algorithms for alternating least-squares CP and Tensor-Train decomposition based on randomized leverage score sampling. Given a fixed accuracy threshold per iteration, our algorithms achieve asymptotically lower runtime than recent state-of-the-art methods. Both algorithms rely on a common data structure that enables implicit sampling from a large, structured probability vector in runtime logarithmic in its length. Experiments on sparse tensors with billions of nonzero entries confirm that our methods enjoy practical runtime and accuracy advantages over existing algorithms. The work presented is a joint collaboration between researchers at UC Berkeley, Lawrence Berkeley National Lab, Sorbonne Université, the International Computer Science Institute, and the University of Montreal.

Day 2 19-October-2023, Session 2, Time: 10:30-Noon EST

Title: Faster Structured Tensor Decompositions via Sketching
Speaker: Alex Gittens, Rensselaer Polytechnic Institute, gittea@rpi.edu
Abstract: Structured tensor decompositions are useful in a variety of fields, including signal processing, image analysis, neuroscience, and machine learning. We encourage structures– sparsity, positivity, etc– by regularizing the factor matrices, and consider empirically and theoretically the use of sketching to trade-off between accuracy and computational cost in the computation of structured CP and Tucker decompositions.

Title: Distributed Large-Scale All-at-Once Count Tensor Decomposition
Speaker: Teresa Ranadive, Laboratory for Physical Sciences, tranadive@lps.umd.edu

Abstract: CANDECOMP/PARAFAC (CP) tensor decomposition is a commonly utilized approach for data exploration. Highly sophisticated algorithms are required to ensure the decomposition accuracy of tensors formed from increasingly large and elaborate data sets. The algorithm CP-POPT has already demonstrated considerable success in obtaining accurate count tensor decompositions, particularly for tensors related to cybersecurity. In this talk, we consider the distributed memory implementations of both CP-POPT, and the more commonly used count tensor decomposition algorithm, CP-APR. We compare the decomposition accuracy, speed, and scalability of both algorithms. Finally, we observe that by warm-starting CP-APR with the output of several iterations of CP-POPT we can quickly obtain accurate decompositions of large-scale count tensors.

Title: Streaming Generalized Canonical Polyadic Tensor Decompositions
Speaker: Eric Phipps, Sandia National Laboratories, etphipp@sandia.gov
Abstract: In this paper, we develop a method which we call OnlineGCP for computing the Generalized Canonical Polyadic (GCP) tensor decomposition of streaming data. GCP differs from traditional canonical polyadic (CP) tensor decompositions as it allows for arbitrary objective functions which the CP model attempts to minimize. This approach can provide better fits and more interpretable models when the observed tensor data is strongly non-Gaussian. In the streaming case, tensor data is gradually observed over time and the algorithm must incrementally update a GCP factorization with limited access to prior data. In this work, we extend the GCP formalism to the streaming context by deriving a GCP optimization problem to be solved as new tensor data is observed, formulate a tunable history term to balance reconstruction of recently observed data with data observed in the past, develop a scalable solution strategy based on segregated solves using stochastic gradient descent methods, describe a software implementation that provides performance and portability to contemporary CPU and GPU architectures and demonstrate the utility and performance of the approach and software on several synthetic and real tensor data sets.

Title: Tensor Train Approach to PDE-Constrained Optimization under Uncertainty
Speaker: Akwum Owunta, Lehigh University
Abstract: This talk discusses a new approach to solving high-dimensional risk-averse optimization problems governed by differential equations (ODEs and/or PDEs) under uncertainty. We focus on the so-called Conditional Value at Risk (CVaR), but the approach equally applies to other coherent risk measures. Both the full and reduced space formulations are considered. Our proposed algorithm is based on low-rank tensor approximations of random fields discretized using stochastic collocation. To avoid the non-smoothness of the objective function underpinning the CVaR, we propose an adaptive strategy to select the width parameter of the smoothed CVaR to balance the smoothing and tensor approximation errors. Moreover, an unbiased Monte Carlo CVaR estimate can be computed using the smoothed CVaR as a control variate. The numerical experiments demonstrate that the proposed method enables accurate CVaR optimization constrained by large-scale discretized systems.

Day 2 19-October-2023, Session 3, Time: 1:30-3pm EST

Title: Convenient development of general tensor network algorithms with ITensor
Speaker: Matt Fishman, Flatiron Institute, mfishman@flatironinstitute.org

Abstract: ITensor is a library for running and developing tensor network algorithms, a set of algorithms where high order tensors are represented as a network of lower order, and low rank, tensors. Originally written in C++, it has since been ported to the Julia language, where it has grown beyond MPS methods like DMRG with recent and ongoing expansions into new areas like quantum circuit simulation using generic tensor network architectures, modern dynamics and optimization algorithms on tree tensor networks, automatic differentiation, and growing support for running dense and block sparse operations on various GPU backends. I will give an update on the ITensor Julia ecosystem. In particular, I will discuss updates to our tensor operation backend library, NDTensors.jl, and efforts to make it more extensible and support dense and block sparse operations on a variety of GPU backends through a simple code design and interface. In addition, I will discuss the new ITensorNetworks.jl library, a library built on top of ITensor for tensor network algorithms on general graphs, which has a graph-inspired interface and support for a variety of graph operations to make it easier to develop new tensor network algorithms.

Title: Automatically Translating Sparse Codes

Speaker: Avery Laird, University of Toronto, laird.avery@gmail.com

Abstract: Optimized libraries and domain-specific languages (DSLs) enable high performance for a diverse range of applications, including sparse tensor kernels. However, manually rewriting existing sparse codes to use libraries or DSLs is tedious, error-prone, and sacrifices portability. Ideally, a tool will translate programs automatically. Existing automatic translation techniques focus on dense, statically analyzable programs. Sparse codes are challenging to analyze statically because they use storage formats, which introduce implicit assumptions about input data. As a result, sparse codes have highly data-dependent semantics that can only be known at run-time. This talk presents REV, a tool to translate data-dependent programs such as sparse codes automatically. REV combines bisimulation techniques with dynamic run-time checks to prove the semantics of sparse codes and guarantee correct translations. An experimental evaluation shows up to 10.4x speedup on a diverse range of benchmarks, with minimal compilation overhead. Additionally, REV's run-time checks detect two bugs in widely-used benchmarks.

Title: Tensor Butterfly Factorization (In Parallel!)

Speaker: Paul Kielstra, UC Berkeley, pmkielstra@berkeley.edu

Abstract: First developed in the context of the Fast Fourier Transform, butterfly factorizations have found applications to the many other matrices that satisfy the complementary low-rank property (CLR). We extend butterfly methods to tensors in four or more dimensions that satisfy an analogous property, achieving both lower compression times and better memory use compared to the usual method of flattening the tensor into a matrix and butterfly-factoring that. The process also suggests a somewhat novel understanding of butterfly factorization itself, allowing us to write algorithms that can be described extremely tersely and parallelize well.

Title: Optimizing Equivariant Tensor Products — the Computational Bottleneck of Symmetry-Equivariant Neural Networks

Speaker: Mit Kotak, Massachusetts Institute of Technology,

Abstract: 3D Geometry (e.g. point clouds, volumes, meshes or surfaces) and geometric tensors (e.g. vector fields or strain fields) are traditionally challenging data types to use for ML applications because coordinates and coordinate systems are sensitive to the symmetries of 3D space: 3D rotations, translations, and inversion. 3D Euclidean symmetry-equivariant neural networks (E(3)NNs) are emerging as an effective machine learning paradigm for processing geometrical information in molecular modeling, protein design, computer graphics, and beyond. E(3)NNs differ from traditional neural networks in three key ways: they have specialized data-types (irreps), multiplication operations (tensor products + decomposition), and means of parameterizing functions of geometry (e.g. convolutional kernels based on basis functions of the group). Currently, symmetry-equivariant neural networks are more computationally expensive than non-equivariant counterparts. However, this is largely an engineering issue. Modern deep learning frameworks like PyTorch and JAX are not optimized for the specialized operations of E(3)NNs. In this talk, I will focus on the tensor product + decomposition operation which increases significantly in computational complexity as higher-order tensors are used. I'll discuss what makes current implementations inefficient as well as possible strategies and early proof of principle examples that may help expedite these operations.