Call for Participation in

Themed session **“Machine Learning and Data Storage”**

at *IEEE Information Theory Workshop (ITW) 2022*
(During online sessions, November 1-2, 2022)

Organizers:

Paul H. Siegel, Distinguished Professor, UC San Diego, psiegel@eng.ucsd.edu

Netanel Raviv, Assistant Professor, Washington University in St. Louis, netanel.raviv@wustl.edu

Anxiao (Andrew) Jiang, Professor. Texas A&M University, ajiang@cse.tamu.edu

Importance and Timeliness:

The current wave of the AI revolution is having a deep impact on data storage systems. On one side, the vast amount of data and models which needed to be stored for machine learning grows exponentially fast. On the other side, machine learning is fundamentally affecting the study of data storage systems, including the modeling of storage channels, compression and privacy of data, error-correction code designs, etc. So the integration of machine learning with data storage offers enormous opportunities for research. At the same time, with hundreds of billions of new machine learning models being stored in devices annually — from mobile devices to cloud servers — the need to explore new algorithms that combine information theory with machine learning for data storage is also highly urgent.

The synergy between machine learning and data storage points to significant mutual benefits. On one hand, new machine learning algorithms can substantially enhance storage systems. Machine learning can be used to model critical information about storage devices, such as channel models and error prediction of flash memories, memristors, etc. It can be used to design new error-correction codes, including for both classic storage channels (BSC, AWGN, etc.) and non-conventional channels (multiple-level memory cells, intercell interference, DNA storage, etc.). It can be used to effectively compress data (e.g., images, videos) at various rate-distortion tradeoffs, and reduce the storage footprint in storage systems. It can recover data from errors with unprecedented capabilities (e.g., through denoising and inpainting for images). It can also be used to protect information privacy, and allow storage devices to be accessed more securely.

On the other hand, information theory is also crucial for machine learning models and their data stored in storage systems. Accompanying the revolution of AI, the sizes of machine-learning models (especially deep learning models) are also growing fast: from models of tens to hundreds of megabytes for computer vision to models of up to hundreds of gigabytes for natural language processing. The data that accompany the models (e.g., for training, testing or inference) are usually even larger. The machine learning models and data all need to be stored in storage systems reliably. As machine learning models are for specific tasks, they open up a new frontier for data storage, which is Application-Specific Data Storage (ASDS) instead of general-purpose data storage (analogous to ASIC versus general-purpose integrated circuits). In particular, how to protect the vast amount of parameters in machine-learning models and their associated data adaptively from errors using error-correction schemes — with the

objective of balancing redundancy and the performance of machine-learning models and optimizing this tradeoff — offers both new challenges and new opportunities. In addition, coding theory can enhance the robustness of common computations in machine learning — including tensor multiplication, gradient descent, etc. — in noisy environments, including low-power computation in mobile devices and collaborative learning in cloud workstations. Machine-learning oriented storage shifts storage systems closer than ever to computation.

Plan of Activities:

The themed session "Machine Learning and Data Storage" will be held virtually over the Internet, following all guidelines of ITW 2022. It will feature five invited talks, each covering a distinct and important topic in the area. Each talk will take 16 minutes, including 14 minutes of presentation and 2 minutes for answering questions from the audience.

A special 10-minute "**From The Audience**" program will be organized to promote broad participation in the themed session, improve interactions among all attendees, and help build a stronger community on "machine learning and data storage". In the 10-minute program, each researcher from the audience can freely and briefly introduce their works related to machine learning and data storage (e.g., who they are, the background of their work, their published papers, etc.). They can also submit their related works to the organizers in advance by email, and the organizers will prepare a succinct presentation to summarize those interesting topics. The program will be moderated by the organizers. The organizers will prepare enough materials to ensure a successful and informative program.

If you are interested in the "**From the Audience**" program and have works related to the theme of this session, you are highly encouraged to email the organizers in advance with an introduction to your work. (The email addresses can be found at the top of this document.)

Presentations:

Presenter: Simeng Zheng (University of California, San Diego)
Topic: Code-aware Storage Channel Modeling via Machine Learning
Introduction:
With the reduction in device size and the increase in cell bit-density, NAND flash memory suffers from larger inter-cell interference (ICI) and disturbance effects. Constrained coding can mitigate the ICI effects by avoiding problematic error-prone patterns, but designing powerful constrained codes requires a comprehensive understanding of the flash memory channel. Recently, we proposed a modeling approach using conditional generative networks to accurately capture the spatio-temporal characteristics of the read signals produced by arrays of flash memory cells under program/erase (P/E) cycling. In this work, we introduce a novel machine learning framework for extending the generative modeling approach to the coded storage channel. To reduce the experimental overhead associated with collecting extensive measurements from constrained program/read data, we train the generative models via transferring knowledge from models pre-trained with pseudo-random data. This technique can accelerate the training process and improve model accuracy in reconstructing the read voltages induced by constrained input data throughout the flash memory lifetime. We analyze the quality of the model by comparing flash page bit error rates (BERs) derived from the generated and measured read voltage distributions. We envision that this machine learning framework will serve as a valuable tool in flash memory channel modeling to aid the design of stronger and more efficient coding schemes.

Presenter: Ron M. Roth (Technion)
Topic: Fault-Tolerant Neuromorphic Computing on Nanoscale Crossbar Architectures
Introduction:
In this talk, recent coding techniques are reviewed for protecting nano-scale crossbar architectures against faults and computational errors. Two computational paradigms are considered: exact computation over the integers, and approximate computation over the reals.

Presenter: Anxiao (Andrew) Jiang (Texas A&M University)
Topic: Symbolic Regression for Data Storage with Side Information
Introduction:
There are various ways to use machine learning to improve data storage techniques. In this talk, we introduce symbolic regression, a machine-learning method for recovering the symbolic form of a function from its samples. We present a new symbolic regression scheme that utilizes side information for higher accuracy and speed in function recovery. The scheme enhances latest results on symbolic regression that were based on recurrent neural networks and genetic programming. The scheme is tested on a new benchmark of functions for data storage.

Presenter: Rawad Bitar (Technical University of Munich)
Topic: Efficient Private Storage of Sparse Machine Learning Data
Introduction:
We consider the problem of maintaining sparsity in private distributed storage of confidential machine learning data. In many applications, e.g., face recognition, the data processed in machine learning algorithms is represented by sparse matrices which can be stored and processed efficiently. However, mechanisms maintaining perfect information-theoretic privacy require encoding the sparse matrices into randomized dense matrices. It has been shown that, under some restrictions on the storage nodes, sparsity can be maintained at the expense of relaxing the perfect information-theoretic privacy requirement, i.e., allowing some information leakage. In this work, we lift the restrictions imposed on the storage nodes and show that there exists a trade-off between sparsity and the achievable privacy guarantees. We focus on the setting of non-colluding nodes and construct a coding scheme that encodes the sparse input matrices into matrices with the desired sparsity level while limiting the information leakage.

Presenter: Eitan Yaakobi (Technion)
Topic: The Zero Cubes Free and Cubes Unique Multidimensional Constraints
Introduction:
This work studies two families of constraints for two-dimensional and multidimensional arrays. The first family requires that a multidimensional array will not contain a cube of zeros of some fixed size and the second constraint imposes that there will not be two identical cubes of a given size in the array. These constraints are natural extensions of their one-dimensional counterpart that have been rigorously studied recently. For both of these constraints we present conditions of the size of the cube for which the asymptotic rate of the set of valid arrays approaches 1 as well as conditions for the redundancy to be at most a single symbol. For the first family we present an efficient encoding algorithm that uses a single symbol to encode arbitrary information into a valid array and for the second family we present a similar encoder for the two-dimensional case. The results in the paper are also extended to similar constraints where the sub-array is not necessarily a cube, but a box of arbitrary dimensions and only its volume is bounded.