Research Experience for Undergraduates

Summer School on Mathematical Foundation of Data Science

June 5, 2023 --- July 14, 2023

LeConte 102

Department of Mathematics

University of South Carolina

Or Join Virtual Zoom Program

https://us06web.zoom.us/j/87902030404?pwd=UVVndERJK2ExS09rWFhYSW5MSk42QT09

Meeting ID: 879 0203 0404
Passcode: 188110

Sponsored by

Department of Mathematics
University of South Carolina

National Sciences Foundation
RTG award DMS 2038080

UNIVERSITY OF South Carolina

Organized by

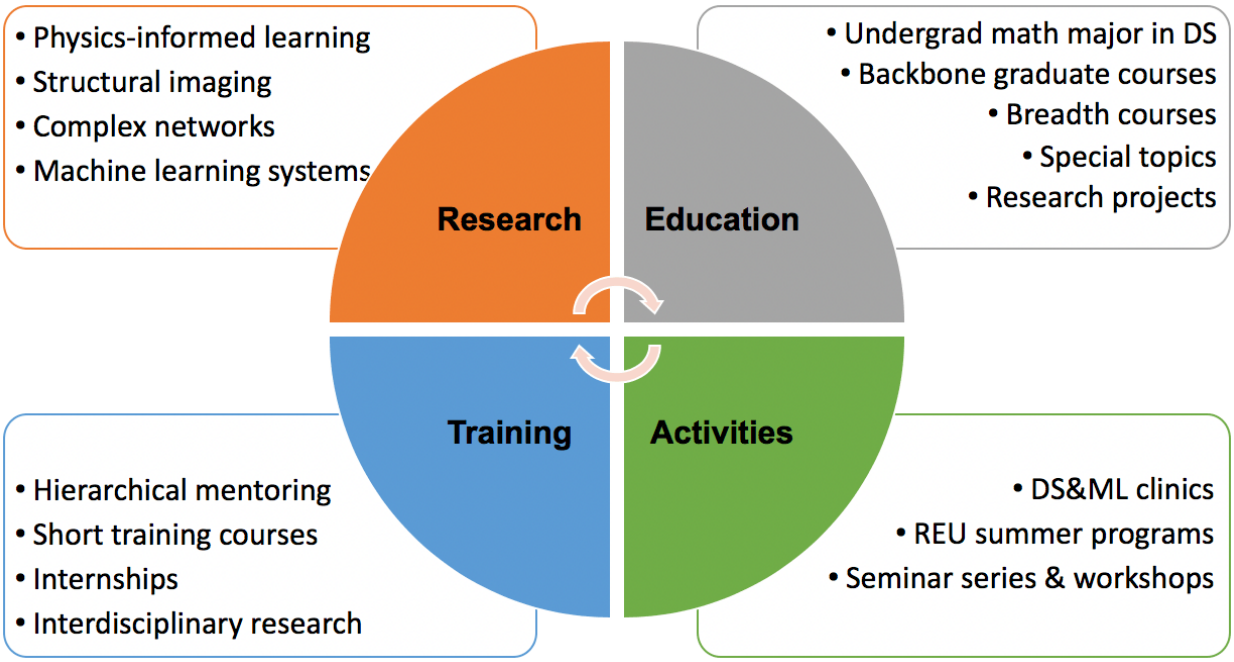
Prof. Linyuan Lu,
Prof. Wuchen Li,
Prof. Qi Wang,
Prof. Changhui Tan,
Prof. Peter Binev,
Prof. Xinfeng Liu

NSF

Table of Contents

Section 1: Program Overview

This REU summer program is part of the NSF RTG project **"RTG: Mathematical Foundation of Data Science at University of South Carolina"**, which aims to develop a multi-tier Research Training Program at the University of South Carolina (UofSC) designed to prepare the future workforce in a multidisciplinary paradigm of modern data science. The education and training models will leverage knowledge and experience already existing among the faculty and bring in new talent to foster mathematical data science expertise and research portfolios through a vertical integration of post-doctoral research associates, graduate students, undergraduate students, and advanced high school students. A primary focus of this project is to recruit and train U.S. Citizens, females, and underrepresented minority (URM) among undergraduate and graduate students, and postdocs through research led training in Data Science.



For more information on the NSF RTG project, please visit us at the following URL: https://sc.edu/study/colleges_schools/artsandsciences/mathematics/my_mathematics/rtg/index.php

The REU summer program of this year runs in-person and virtually from June 6 to July 15. Starting from the first week, students will be divided into serveral groups to work on research projects. Some guest speakers are invited to give talks on the latest development in the Mathematical Foundation of Data Science. On the last day of the program, students will present their research findings.

Section 2: Research Projects

2.1 Research project on subdivision schemes for processing multi-modal data (Advisor: Peter Binev)

There are several practical problems in which the data about the investigated specimen comes in different modes. For example, in Atomic Force Microscopy (AFM) there are two modes: tapping and contact modes. The tapping mode utilizes a vertical oscillation of the probe over the surface of the specimen and allows creation of high-resolution images of the surface. In contact mode the tip is in constant contact with the surface causing shearing forces and damage to the surface or the tip resulting in lower resolution images. The main question is how to increase the resolution of the data received in the contact mode. Stationary subdivision schemes are often used to predict the higher resolution versions of smooth curves and surfaces. However, the AFM data is not smooth and to make a proper prediction one needs to change locally the schemes, making them nonstationary. The high-resolution tapping mode can provide information about the appropriate changes. The research project aims at investigating different variants of nonstationary subdivision schemes and the ways of determining their coefficients based on the available high-resolution data for the specimen. Steps in the project:

1. In one dimension, investigate different variants of determining of the local coefficients of predictive nonstationary subdivision schemes using the data mode with known high resolution.
2. Apply the above variants on lines of the real data and analyze their relevance.
3. Design a two-dimensional predictive nonstationary subdivision scheme that consistently produces reliable results for the available data. Test it on new data.

2.2 Generating Synthetic STM Data for Training of DNNs (Advisor: Peter Binev)

The training of Deep Neural Networks (DNNs) usually requires enormous amounts of data. In many cases the available data is not enough, it is expensive to generate, and often difficult to label properly. The solution is to generate data using a mathematical model that describes well the investigated phenomena and can produce synthetic data that closely resembles the real one especially regarding the features of interest.

The aim of this project is to generate training data representing the results from Scanning Tunneling Microscopy (STM), see https://en.wikipedia.org/wiki/Scanning_tunneling_microscope. STM gives information about the atomic structure on the surface of the investigated material. Each data file is an image (2D array of positive values) in which the atoms appear as round spots assembled in different patterns representing the local structure of the materials. Steps in the project:

1. For different 3D Bravais lattices, starting with the cubic ones, calculate the 2D patterns corresponding to tilted plane cuts through the lattice in which the atom spots are calculated based on the lattice points closest to the plane with size and intensity depending on the distance to the plane.

2. Assemble an STM image combining two or more patterns in separate areas of the image using different conventions around the boundaries between the areas inspired by real images.

3. Add noise and distortions to the synthetic STM images to make them more realistic.

2.3 Research projects in transport information learning and optimization (Advisor: Wuchen Li)

Study and understand natural gradient methods from information geometry and optimal transport. Implement the natural gradient algorithms for supervised learning problems, and unsupervised learning problems.

1. In one-dimensional space, compute and implement the Fisher and Wasserstein information matrix for Gaussian and exponential distributions. Then, implement the natural gradient methods to learn the parameters.

2. In discrete graphical models, compute and implement the Wasserstein natural gradient methods for learning parameters in Boltzmann machines.

3. In two-layer neural network models, compute and implement the Wasserstein information matrix and its induced natural gradient dynamics.

2.4 Research project in data-driven mathematical modeling of cancer tumor growth (Advisor: Xinfeng Liu)

Breast cancer is a malignant disease with a heterogeneous distribution of cell types. Cancer stem cells (CSCs) are defined as "a small subset of cancer cells" within a cancer that can self-renew and replenish the heterogeneous lineage of cancer cells that comprise the tumor. Oncogene HER2 is a well-studied oncogene known for its role in promoting cancer cell survival and proliferation, and recently implicated in generation and maintenance of CSCs. With a close integration with experimental data, the main theme of this project is on the mathematical modeling and computational investigation of Cancer Stem Cell (CSC) population in cancer tumor growth coupled with Her2 signaling pathways.

1. In experimental data fitting, apply and implement optimization to characterize the cancer cell population growth.
2. In mathematical modeling, develop and evaluate various mathematical models to explore underlying mechanisms and biological factors to control the balance of tumor growth.
3. In comparison of mathematical modeling with experimental data, develop and implement numerical algorithms for searching parameters with data fitting.

2.5 Research project on nonlocal traffic and pedestrian flows (Advisor: Changhui Tan)

The study of mathematical theory related to traffic and pedestrian flows dates back to the 1920s and has remained an active area of research throughout the last century. Numerous models have been proposed and studied, aiming to comprehend the intricate mechanisms that lead to traffic congestion. The development of this theory has significantly contributed to the improvement of designing efficient transportation systems for traffic and pedestrian networks.

This project aims to explore various models pertaining to traffic and pedestrian flows, with a particular focus on utilizing global information to optimize network efficiency.

1. Investigate classical models in traffic and pedestrian flows, and introduce nonlocal interactions as a means of global control within the models.
2. Devise numerical methods and simulate both local and nonlocal models. Investigate the mechanisms influenced by the introduction of nonlocal interactions.
3. Integrate real data into the research and design algorithms for learning the nonlocal interactions.

By undertaking these steps, we aim to gain a deeper understanding of nonlocal traffic and pedestrian flows, and explore novel approaches for enhancing network efficiency based on global information.

2.6 Reciprocal regulation of breast tumor growth and systemic metabolism in mice (Advisor Qi Wang)

We develop a predictive model for identifying prominent features of cancer-laden mice by contrasting healthy mice with cancer-laden ones in terms of a host of behavioral and metabolic biomarkers. Some of the data are collected as time series every 30 mins while others are static data.

Introduction:

Rationale: Tumor growth may impact host metabolism, and host metabolism likely modulates tumor development.

Basic science questions: How does breast tumor growth alter systemic metabolism? Do certain systemic metabolism features affect tumor growth?

Clinical questions: Can certain subtle systemic metabolism (longitudinal) changes be used as biomarkers for early cancer detection or cancer progression monitoring? Can certain metabolism interventions slow down tumor growth?

Design: Total 16 mice were used. 4 as healthy control (no tumor inoculated); 12 mice were inoculated with breast cancer cells in the mammary fat pad. After cancer cell inoculation, all mice were individually

placed in metabolic cages for the measurement of metabolism-related parameters for 24 days. The mice were euthanized on Day 24 post-cancer cell inoculation. Blood samples were collected for blood glucose measurement. Tumors (2) in each mouse were measured and sizes were averaged.

Methods:

We first preprocess the data and then apply order reduction to represent the data of each mouse in the given dataset in lower dimensional latent space. This data preprocessing will be accomplished by PCA and VAE. The resulting dataset will consist of longitudinal data with fewer variables/markers. In the reduced-order datasets of all mice, we will analyze their correlational relationships, dynamical patterns and potential causal relationships.

2.7 Research problems in Graph Theory and Network Science (Advisor: Linyuan Lu)

Graph theory has emerged as a primary tool for detecting numerous hidden structures in various information networks, including Internet graphs, social networks, biological networks, or, more generally, any graph representing relations in massive data sets. We will focus on some research problems related to spectral graph theory and fractional chromatic number.

Problem 1: Consider a graph G on n vertices. There are several ways to associate a matrix to a graph G. The eigenvalues of these matrices characterize important properties of graphs. For example, the Laplacian matrix L is defined as D-A, where D is the diagonal matrix of degrees of G and A is the adjacency matrix of G. The second smallest eigenvalue of L is called the algebraic connectivity of G.

Which planar graph on n vertices has the maximum algebraic connectivity?

Problem 2: Consider a graph G on n vertices with adjacency matrix A. The largest eigenvalue of A (called the spectral radius) and the smallest eigenvalue have been intensively studied. It is known, for example, that complete graphs have maximum largest eigenvalue. Conversely, much less is known about the "middle eigenvalues" of A. For example, it is not known which graph maximizes the third largest eigenvalue of the graph.

The question is: how to improve on the known bounds and constructions for maximizing the "middle" eigenvalues of the graph? Similar questions can be explored for other matrices associated to a graph or combinations of eigenvalues.

Problem 3: The famous Brooks' Theorem states that if the maximum degree of a connected graph G is Δ, then the chromatic number χ(G) ≤ Δ unless G is a complete graph or an odd cycle. The fractional analogue of Brook's theorem states $\chi_f(G) < \Delta$ that unless G is a complete graph, an odd cycle, C_8^2 , or C5 ⊠ K2. The question is to determine best constant c_Δ so that $\chi_f \leq \Delta - c_\Delta$ for all other graphs.

Dvorak, Sereni and Volec proved $\chi_f \leq \frac{14}{5}$ if G has maximum degree at most 3. The problem is widely open for $\Delta \geq 4$.

Section 3: Program Calendar

The calendar is subject to be revised as needed.

Week 1 (Week of June 5-9)

Day		Activity	Instructor/ moderator
Monday June 5	9:00-12:00	Introduction to the REU program	
	12:00-2:00	Lunch break	
	2:00-5:00	Group assignments	
Tuesday June 6	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Wednesday June 7	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Thursday June 8	9:00-11:00	Zoom Meeting: Research Projects in Computational Microscopy	Peter Binev
	11:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Friday June 9	9:00-12:00	Parallel and joint research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Social/free time	

Week 2 (Week of June 12-16)

Day		Activity	Instructor/ moderator
Monday June 12	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Tuesday June 13	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Wednesday June 14	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	

Thursday June 15	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Friday June 16	9:00-12:00	Parallel and joint research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Social/free time	

Week 3 (Week of June 19-23)

Day		Activity	Instructor/ moderator
Monday June 19	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Tuesday June 20	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Wednesday June 21	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Thursday June 22	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Friday June 23	9:00-12:00	Parallel and joint research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Social/free time	

Week 4 (Week of June 26- June 30):

Day		Activity	Instructor/ moderator
Monday June 26	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Tuesday	9:00-12:00	Parallel research sessions	

June 27	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Wednesday June 28	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Thursday June 29	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Friday June 30	9:00-12:00	Parallel and joint research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Social/free time	

Week 5 (Week of July 3-7)

Day		Activity	Instructor/ moderator
Monday July 3	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Tuesday July 4		Holiday, no activity.	
Wednesday July 5	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Thursday July 6	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Friday July 7	9:00-12:00	Parallel and joint research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Social/free time	

Week 6 (Week of July 10-14):

Day		Activity	Instructor/ moderator
Monday	9:00-12:00	Parallel research sessions	

July 10	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Tuesday July 11	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Wednesday July 12	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Thursday July 13	9:00-12:00	Parallel research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Parallel research sessions	
Friday July 14	9:00-12:00	Parallel and joint research sessions	
	12:00-2:00	Lunch break	
	2:00-5:00	Group Presentations	

Section 4: Participants contact information

Faculty

Linyuan Lu	Lu@math.sc.edu	https://people.math.sc.edu/lu/
Peter Binev	Binev@math.sc.edu	https://people.math.sc.edu/binev
Wuchen Li	wuchen@mailbox.sc.edu	https://people.math.sc.edu/wuchen/
Xinfeng Liu	Xfliu@math.sc.edu	https://people.math.sc.edu/xfliu/
Changhui Tan	tan@math.sc.edu	https://people.math.sc.edu/tan/
Qi Wang	qwang@math.sc.edu	https://people.math.sc.edu/wangq/

Postdoctoral researcher

William Linz	wlinz@mailbox.sc.edu

Graduate assistants

George Brooks	GHBROOKS@email.sc.edu
Black McKenzie	MMBLACK@email.sc.edu
Sehwan Kim	Sehwank@email.sc.edu

Undergraduate students

Cadence	Masciotro	In-Person	masciotc@email.sc.edu	University of South Carolina
Jack	Hyatt	In-Person	Jahyatt@email.sc.edu	University of South Carolina
Connor	Chonko	In-Person	cchonko@email.sc.edu	University of South Carolina
Rebecca	Wang	In-Person	rebeccawang@gatech.edu	Georgia Institute of Technology
Maggie	Gu	In-Person	meiqi.gu@gmail.com	University of North Carolina
Aaron	Wu	Virtually	aw80@rice.edu	Rice University
Philip	Howe	In-Person	pchowe@email.sc.edu	University of South Carolina
Matthew	Belanger	In-Person	belangem@email.sc.edu	University of South Carolina
Olivia	Hoyhtya	Virtually	ohoyhtya@email.sc.edu	University of South Carolina
Ryan	Oscherwitz	Virtually	oscherwitz1@kenyon.edu	Kenyon College
Riley	Gleaton	Virtually	gleator@email.sc.edu	University of South Carolina
Zoe	Shive	Virtually	zshive@email.sc.edu	University of South Carolina
Grace	Wang	Virtually	gswang@ucsd.edu	University of California, San Diego
Niklas	Chang	Virtually	nhchang@ucsd.edu	University of California, San Diego
Justin	Che	Virtually	Jche@wpi.edu	Worcester Polytechnic Institute
Kevin	Hu	Virtually	kevindhu683@gmail.com	Georgia State University
Alexander	Yuan	Virtually	alexanderyuan12@gmail.com	California Polytechnic University
Zach	Li	Virtually	lizach739@gmail.com	The University of Texas at Austin
Shawn	Fong	Virtually	f.shawn@wustl.edu	Washington University in St. Louis