The UC Davis TETRAPODS Institute of Data Science (UCD4IDS)

Naoki Saito

Department of Mathematics, University of California, Davis

The 3rd NSF TRIPODS PI Meeting, March 12, 2020, Atlanta, Georgia

Structure

- Composed of 4 PIs, 31 senior participants, 3 postdocs, 15 graduate students; coming from **CS**, **ECE**, **Math**, **Stat**
- Crosses interdepartmental barriers and promotes interdisciplinary research collaborations among faculty, postdocs, and graduate students
- Collaborates with the other internal data science groups, e.g., the *CeDAR* (Center for Data science and Artificial intelligence Research) that includes domain experts from medicine, agriculture, astronomy, etc., and the *DataLab* that provides support and training

Research Themes

I: Fundamentals of machine learning directed toward biological and medical applications

II: Optimization theory and algorithms for machine learning including numerical solvers for large-scale nontrivial learning problems

III: High-dimensional data analysis on graphs and networks

I: Fundamentals of Machine Learning

- Members:
  CS: N. Amenta, P. Devanbu, P. Koehl, Y.-J. Lee, I. Tagkopoulos;
  ECE: C.-N. Chuah, S. Ghiasi;
  Math: J. Arsuaga, J. De Loera, J. Hass, L. Rademacher, M. Vazquez;
  Stat: C. Drake, F. Hsieh, J. Jiang, M. Lopes, W. Polonik, B. Rajaratnam

Ia: Geometry of Data

- *Morphology of biological systems:* changes of protein structure, tissues, brain structures during their development, aging, learning, disease, and evolution
- *Analysis of genomic data:* via *Topological Data Analysis*
- *Data clustering and classification:* generalization of existing spectral graph clustering methods using the higher-order *Hodge Laplacian*

Ib: Pattern Mining and Machine Learning

- *Novel data acquisition:* *embedded/wearable devices* for biomedical and health applications
- *Data-harness and feature detection/selection:* optimal experimental design and active learning for high throughput biological experiments (RNA-Seq, omics, etc.); computer vision for learning scalable recognition systems
- *Statistical learning in high-dimensional data:* large-scale computation via upgraded *bootstrapping* techniques
- *Unsupervised learning:* extraction of information content for computable knowledge; coherent decision-making with available information
- *Sampling and streaming:* small-area sampling on big data; propensity weighting; column/row sampling of matrices with optimal approximation guarantees; sampling and streaming to capture the statistical properties of cascades in large online social network graphs

II: Optimization for Machine Learning

- Members:
  CS: P. Devanbu, I. Tagkopoulos;
  ECE: L. Lai;
  Math: J. De Loera, A. Fannjiang, M. Köppe, S. Ma, T. Strohmer;
  Stat: K. Balasubramanian, X. Li
- *Stochastic algorithms:* deeper understanding and performance improvement of *Stochastic Gradient Descent* (SGD) algorithm
- *Optimization landscape of nonconvex problems:* investigation on quality of local minima; the issue of *escaping from a saddle point* via randomization/perturbation for SGD and its variants
- *Privacy and security in machine learning:* *on-device* machine learning with limited resources; *distributed* optimization algorithms
- *Applications—Phase retrieval and beyond:* *ptychography* phase retrieval with nonlinearly coupled coded diffraction patterns

III: Data Analysis on Graphs & Networks

- Members:
  CS: P. Devanbu, I. Tagkopoulos;
  ECE: C.-N. Chuah, Z. Ding;
  Math: R. Chaudhuri, N. Saito;
  Stat: A. Aue, K. Balasubramanian, S. Chen, T. Lee, C. Le, X. Li, D. Paul, J. Sharpnack
- *Hypergraphs and tensors:* nonparametric modeling of *higher-order interactions* between nodes; non-negative tensor decomposition algorithms; extension of the *Generalized Haar-Walsh Transforms* from graphs to hypergraphs
- *Collective computation in distributed neural systems:* identification of the theoretical frameworks and dynamical principles underlying parallel distributed computation in neural systems; modeling the plasticity rules of neurons; extraction of neural coding from neural activities
- *High-volume neural data analysis:* investigation of point processes with history-dependent intensity for spike train data analysis; inference of the neural microcircuits using optogenetic stimulation
- *Anomaly and change detections over networks:* computationally efficient and statistically optimal change detectors of graph data and graph structure
- *Random matrix theory for network applications:* extraction of latent graph structures from high-dimensional time series (e.g., fMRI, EEG datasets); modeling and prediction of potentially non-Gaussian stochastic processes on graphs; clustering sparse random networks and understanding associated concentration phenomena
- *Network adaptation for machine learning:* design of *ML-aware networks* that maintain a certain level of image classification accuracy while minimizing the bandwidth consumption through optimizing quantization parameters for transformed coefficients of images

Educational Activities

- Some members have been involved in a campus-wide committee on *Undergraduate Data Science Degree Program*:
  - Will have "Core" DS Track (focusing on Math/Stat/CS); Bio DS Track; Ag-Environment DS Track; Humanities/Social Sci DS Track
  - Everyone needs to take a set of core courses in Math/Stat/CS, which are in the process of approval
  - Then, students in each track take track-dependent courses
- At the graduate level, we collected all the courses relevant for data science research, and put their links on our website
- Will initiate a procedure to establish the *Graduate Group in Data Science* in a few years

Community Building Activities

- Developed *our website* `https://ucd4ids.ucdavis.edu` as a vehicle for centralizing our data science related information
- *Weekly seminars:* Mathematics of Data and Decision Seminar (Math); Statistics Seminar
- *Quarterly colloquia* (including Joint Math/Stat Colloquium)
- *Roundtable discussions* associated with the above seminars, whose *minutes* are disseminated through our website
- Hiring *three postdocs* who will play important roles in our projects