ZHEN SHAO

Mobile: +447849407273, Email: shaoz@maths.ox.ac.uk

Education

PhD in Mathematics, University of Oxford **2017 – 2022**

• Thesis: On Random Embeddings and Their Application to Optimisation. Supervisor: Prof Coralia Cartis. Submitted, January 2022 and passed viva with minor corrections in March 2022.
• Awarded full scholarship (£140,000+).

Master of Mathematical and Theoretical Physics, University of Oxford **2016 - 2017**

• Distinction, top three of year group (70 students).
• Mini-project: Power and centrality in networks, a first-principles approach.

Bachelor of Mathematics, University of Oxford **2013 - 2016**

• First class, top 5% of year group (200 students).
• Summer project: Mathematical methods for website optimisations.
• Took double the number of required modules.

Work Experiences

Randomness in Machine Learning Algorithms **2022 - Current**

Postdoctoral Researcher, Mathematical Institute, University of Oxford, Oxford

• Aim: To further the understanding of the role and benefits of randomness in improving the scalability of algorithms and models in machine learning and optimization, by taking further and expanding the directions pursued in my PhD project.
• Funded by the EPSRC Doctoral prize (six months).
• Result: Ongoing. So far have written results concerning random sampling/projections for noisy data inputs. Currently pursuing theoretical understanding of the relationship between randomisation and generalisation errors of learning algorithms.

Teaching Experiences

Pembroke College, University of Oxford **2022 - 2023**

• Lecturer teaching Probability, Statistics and Data Analysis, Numerical Analysis, Calculus of Variation, Graph Theory, Constructive Mathematics, Introductory Calculus.

Pembroke College, University of Oxford **2021 - 2022**

• Lecturer teaching Differential Equations 1, Differential Equations 2, Quantum Theory, Special Relativity, Waves and Fluids, Introductory Calculus.

Mathematical Institute, University of Oxford **2018 – 2022**

• Class Tutor in Probability, Measure and Martingales
• Class Tutor in Continuous Optimisation
• Tutor in Statistics, Regent's Park college
• Tutor in Numerical Analysis, St Hugh's college
• Teaching assistant in Continuous Optimization
• Departmental MATLAB demonstrator
• Teaching assistant in Integer Programming

Research Publications

• *Johnson-Lindenstrauss embedding for noisy vectors, taking advantage of the noise.* ***Z. Shao****, arXiv preprint, 2022.* https://arxiv.org/abs/2209.01006

• *Randomised subspace methods for non-convex optimization, with applications to nonlinear least-squares.* ***Z. Shao****, C. Cartis and J. Fowkes. Submitted to IMA Journal in Numerical Analysis, 2022. To be made available on arXiv preprint.*
• *On Random Embeddings and Their Applications to Optimisation.* ***Z. Shao****, DPhil Thesis, examined by Professor Michael Mahoney and Professor Raphael Hauser, 2022*. https://arxiv.org/abs/2206.03371
• *Hashing embeddings of optimal dimension, with applications to linear least squares. C. Cartis, J. Fiala and* ***Z. Shao,*** *arXiv preprint, 2021.* https://arxiv.org/abs/2105.11815
• *Sparse sketching for sparse linear least squares. C. Cartis, J. Fiala and* ***Z. Shao****, In Thirty-seventh International Conference on Machine Learning, 2020. In Workshop on Beyond First Order Methods in ML systems.*
https://drive.google.com/file/d/1BacyZwtZSKZBBblLC7x4SikuDQaLHuYK/view
• *A Randomised Subspace Gauss-Newton Method for Nonlinear Least-Squares. C. Cartis, J. Fowkes and* ***Z. Shao****, In Thirty-seventh International Conference on Machine Learning, 2020. In Workshop on Beyond First Order Methods in ML systems.*
https://drive.google.com/file/d/14naXqmbHx9p2ONFwvSfbxAB-QLUAbE_R/view
• One journal paper in preparation from my PhD thesis.

Software Publication

• Ski-LLS, a high-performance C/C++ solver for solving large-scale linear least squares.
https://github.com/numericalalgorithmsgroup/Ski-LLS

Programming Skills

• Intermediate in **C**, experiences in **Python**-based projects**,** Proficient in **MATLAB.**
• Experiences in **Bash, Linux, Git/GitHub, Scala, Haskell, OCaml, R.**

Conferences and Talks

• *Random Embeddings and their Applications to Optimisation Problems*. SIAM Optimisation, 2021.
• *Sketching for Linear Least Squares,* ICCOPT 2019
• *Sketching for Sparse Linear Least Squares,* 17th Workshop on Advances in Continuous Optimization 2019
• *Sketching for Sparse Linear Least Squares,* 29th Biennial Numerical Analysis Conference, 2019
• *Drying of Colloid Suspension,* Numerical Analysis Group Internal Seminar, 2018

Computational Project Experiences

Fast Non-linear Least-Squares Solver **2020**
Numerical Algorithm Group Ltd., Oxford
• Designed an algorithm for solving large-scale non-linear regressions using iterative randomisation of the parameter search space.
• Co-implemented a Python code for solving large-scale non-linear regressions, achieving 4-10 times speed-up on certain logistic regression problems.

Fast Linear Least-Squares Solver **2019**
Numerical Algorithm Group, Oxford
• Designed an algorithm for solving large-scale linear least-squares/regressions using random samplings and random projections of the dataset.
• Derived theoretical guarantees of the solution quality generated by the randomised algorithm.
• Implemented Ski-LLS, a C/C++ software package for solving large-scale linear regression problems. The solver runs 10 times faster than state-of-the-art solvers for problems with 120,000 rows and 50,000 columns. The speed improvement is theoretically better for larger problems.

Simulating the Evaporation of Quantum Dot Suspension Droplets **2018**
Sharp Laboratory Europe Ltd., Oxford
• Created a mathematical model for the manufacturing process of QLED displays.
• Wrote a MATLAB software package simulating the manufacturing process to 99.9% accuracy.
• Recommended manufacturing improvements with over 30% potential material savings.

Data Science Project Experiences

Robust Artificial Intelligence **2021**
Quantum Black, a McKinsey Company, London
• Problem: How to make machine learning models more accurate in the following two scenarios: (1) The testing data is similar, but different, to the training data. E.g., Using data on red wines to predict the quality of white wines. (2) The data is a continuously evolving stream with drifts in unknown locations (that is, the distribution of the data changes, e.g., traffic data before and after COVID).
• Method: Worked extensively in a small team with Quantum Black data scientists and engineers.
• Result: Designed, implemented and published a Python library (RAI) that detects the distribution shifts in data and improves the accuracy of a given model. Disseminated findings by holding internal seminars. Proposed a novel algorithm to dynamically detect the set of predictive features in an evolving data stream.

Differential Privacy in Travel Data **2021**
Resource System Group Ltd., Virtual
• Problem: How could one mask data containing sensitive personal information (in this case, the daily trip data for individuals) so that personal information is protected but key information is preserved?
• Method: Reviewed the concept of differential privacy in the literature (a class of algorithms that are not sensitive to small changes of the data); Use the coin flipping techniques to mask the data (randomly change the categorical response variables for each data entry).
• Result: A python package applying the coin flipping techniques to mask dataset.

Robust Convolutional Neural Network Against Adversarial Attack **2019**
Defence Science and Technology Research, UK
• Problem: Exploring defence techniques for neural network-based image classifiers; because such classifiers are known to be sensitive to (small, not human-detectable) perturbations of the data.
• Method: Reducing sensitivity through pre-processing the input (e.g., reduce resolution, bit-depth).
• Result: Using bit-depth reduction increases the neural network accuracy from 50% to 90% when faced with a 20% imagine modification attack.

Online Dialogue Classifier **2018**
National Health Service, UK
• Problem: Exploring helping mental health telephone services with a dialogue classifier that indicates if the conversation needs de-escalating.
• Method: Use three categories, good, neutral, bad to represent the state of the conversation; Selected features from the transcribed conversation using histogram analysis; used the hidden Markov model, and alternatively the Bayes factor method to classify the state of the conversation.
• Result: Achieved a global classification accuracy of 94.7% with AUC=0.98.

Reduce Electricity Peak Demand by Optimally Using a Battery Storage System **2018**
Graduate Modelling Camp, Oxford
• Problem: Large electricity peak demand increases the pressure on the delivery network. How do we charge/discharge batteries with limited capacity to reduce the peak demand?
• Method: Use S-ARIMA/Gaussian processes to predict the future demand using the historical time-series demand. Minimise the peak demand (with batteries charging/discharging) using a least square optimiser.
• Result: Achieved a 7% peak reduction comparing to the original peak demand.