Data Science Down Under Worksop

8-12 December 2019, Newcastle, Australia

Contributed Talks

1. **Vivak Patel, University of Wisconsin – Madison, US**

   Time: Wednesday, 11 Dec 2019, 11:10am-11:40am

   Title: On the Practice of Solving Randomly Sketched Linear Systems

   Abstract: Randomized linear system solvers have become popular as they have the potential to reduce floating point complexity while still achieving desirable convergence rates. One particularly promising class of methods, random sketching solvers, has achieved the best known computational complexity bounds in theory, but is blunted by two practical considerations: there is no clear way of choosing the size of the sketching matrix apriori; and there is a nontrivial storage cost of the projected system. In this work, we make progress towards addressing these issues by implicitly generating the sketched system and solving it simultaneously through an iterative procedure. As a result, we replace the question of the size of the sketching matrix with determining appropriate stopping criteria; we also avoid the costs of explicitly representing the sketched linear system; and our implicit representation also solves the system at the same time, which controls the per-iteration computational costs. Additionally, our approach allows us to generate a connection between random sketching methods and randomized iterative solvers (e.g., randomized Kaczmarz method). As a consequence, we exploit this connection to (1) produce a new convergence theory for such randomized iterative solvers, and (2) improve the rates of convergence of randomized iterative solvers at the expense of a user-determined increases in per-iteration computational and storage costs.

2. **Daniel Ahfock, University of Queensland, Australia**

   Time: Wednesday, 11 Dec 2019, 11:40am-12:10pm

   Title: On Randomised Sketching Algorithms and the Tracy-Widom Law

   Abstract: There is an increasing body of work exploring the integration of random projection into algorithms for numerical linear algebra. The primary motivation is to reduce the overall computational cost of processing large datasets. A suitably chosen random projection can be used to embed the original dataset in a lower-dimensional space such that key properties of the original dataset are retained. These algorithms are often referred to as sketching algorithms, as the projected dataset can be used as a compressed representation of the full dataset. Sketching algorithms offer probabilistic bounds on the discrepancy introduced from using the sketched dataset in place of the full dataset. We show that random matrix theory, in particular the Tracy-Widom law, is useful for describing the operating characteristics of sketching algorithms. We develop asymptotic approximations for the success rate in generating random subspace embeddings and the convergence probability of iterative sketching algorithms using elements of random matrix theory. We test a number of sketching algorithms on real large high-dimensional datasets and find that the asymptotic expressions give accurate predictions of the empirical performance.

3. **Lindon Roberts, Australian National University, Australia**

   Time: Wednesday, 11 Dec 2019, 12:10pm-12:40pm

   Title: Improving the Scalability of Model-based Derivative-free Optimization

   Abstract: Derivative-free optimization (DFO) methods are an important class of optimization routines for many problems in data science, such as hyperparameter optimization and adversarial attacks for neural networks. However, in model-based DFO methods, the computational cost of constructing local models and

Lagrange polynomials can be high. As a result, these algorithms are not as suitable for large-scale problems as derivative-based methods. In this talk, I will introduce a derivative-free method based on exploration of random subspaces, suitable for nonlinear least-squares problems. This method has a substantially reduced computational cost (in terms of linear algebra), while still making progress using few objective evaluations. I will also discuss how this approach may be extended to DFO for general nonlinear optimization problems.

4. **Peter Eades, University of Sydney, Australia**

   Time: Wednesday, 11 Dec 2019, 2:00pm-2:30pm

   Title: Large Graph Visualisation Using Spectral Sparsification

   Abstract: Visual Analytics of large social networks is a key component of business, defense, and social strategies. Similarly, Network Visualisation plays an important role in Software Engineering and Biotech. The main challenge for visualisation in these fields is the scale problem: current networks are too large for traditional algorithms. In this talk we describe progress toward sub-linear-time visualisation methods for graph visualisation. In particular we describe a promising method based on spectral sparsification.

5. **Mikhail Kamalov, INRIA, France**

   Time: Wednesday, 11 Dec 2019, 2:30pm-3:00pm

   Title: Semi-supervised VAE with PageRank

   Abstract: We present a scalable semi-supervised learning (SSL) framework based on the combination of variational autoencoder (VAE) and PageRank algorithm. We employ latent variables to avoid high-dimensional computations. In our framework we propose to calculate a nodes similarity matrix using pairwise Jensen-Shannon divergence between latent variables inferred from VAE. Then we append the modified Laplacian regularization loss with respect to the difference between latent variables from VAE. This way we can apply PageRank to compute the explicit node predictions and inject it into SSL VAE loss with modified Laplacian regularization loss. It means that the iterative improvement of the similarity matrix positively influences classification results from PageRank. This approach allows us to minimize classification loss over all nodes instead of just the labelled nodes as in SSL VAE and to take into account the geometric structure of the data. Our framework outperforms state of the art models in the batch mode on data with graph structure (e.g., cora, citeseer, pubmed) as well as on data without default graph structure as in image datasets (e.g., mnist).

6. **Zdravko Botev, University of New South Wales, Australia**

   Time: Wednesday, 11 Dec 2019, 3:00pm-3:30pm

   Title: How Bayes Can Help Frequentist Model Selection

   Abstract: In this talk we explain how and why current frequentist lasso-type model selection methods can sometimes fail using a few simple stylistic examples. Using Bayesian ideas of regularization, we suggest a simple remedy that, at least in the examples we tested, significantly improved the identification of the correct model. We conclude that, despite more than 25 years of research into frequentist model selection, the methodology is still not mature and can easily yield misleading results.

7. **Scott Lindstrom, Hong Kong Polytechnic University, Hong Kong**

   Time: Wednesday, 11 Dec 2019, 4:00pm-4:30pm

   Title: Splitting Methods for Signal Recovery

   Abstract: Many problems in data science take on the general form of identifying a pattern (signal) with desirable qualities in a set of data. Such problems are usually approached with the better-known tools of

machine learning (e.g. clustering, regression, support vector machines, neural networks) whose theory is generally analysed through statistical methods or variational analysis. Supervised methods assume the existence of some training data with a priori knowledge of correct results (categorical, numerical, or otherwise) that can be used to construct maps (learning). Unsupervised methods search without such a priori knowledge, but they generally only seek to satisfy some threshold ratio of desirability to complexity (e.g. clustering) rather than more complicated constraints that enforce high specificity for the solution. We will address a problem frequently encountered in engineering contexts: searching for a signal when high specificity is needed and only some structural information about the solution is available. Such problems are called "feasibility problems." When we can describe this structural information as a set of mathematical sets, called constraint sets, for which projections or their approximates are computable, we can employ splitting methods to search for the signal. Research into iterated methods for feasibility problems is expanding rapidly, owing to the wide scope and structural variability of problems amenable to solution by them. In this talk, we will introduce the basic problem. We will then discuss the particulars that are worth considering when choosing which method to use. Specifically, we'll discuss convexity vs nonconvexity of constraints, the number of constraints, and problem size. We will illustrate with examples throughout. We will conclude by introducing some new research on intelligent acceleration for such methods.

8. Sevvandi Kandanaarachchi, Monash University, Australia

Time: Wednesday, 11 Dec 2019, 4:30pm-5:00pm

Title: Dimension Reduction for Outlier Detection

Abstract: Sometimes outliers are buried in noise. In such instances, we need to reduce dimensions to effectively identify outliers. In this talk, we introduce a new set of basis vectors called DOBIN designed for outlier detection. DOBIN brings outliers to the forefront making it easier to detect them. We demonstrate DOBIN's effectiveness using an extensive data repository. In addition, we present some interesting examples of outlier visualization using DOBIN.

9. Fred Roosta, University of Queensland, Australia

Time: Thursday, 12 Dec 2019, 9:00am-9:30am

Title: Reproducing Stein Kernel Approach for Correcting Approximate Sampling Algorithms

Abstract: We provide a post-hoc rejection-free method of obtaining a consistent estimator for quantities related to a target distribution of interest by using samples obtained from an ergodic Markov chain with an arbitrary stationary distribution. The approach involves Stein importance sampling, based on minimisation of the kernelized Stein discrepancy, and is shown to be valid under certain conditions on the mixing of the chain. To demonstrate the practical implications of the method, we show these conditions are satisfied for a large number of unadjusted samplers. We conduct a numerical study showing that the method is super-efficient in practical scenarios that can involve both a large number of parameters and a large data set.

10. Ali Eshragh, University of Newcastle, Australia

Time: Thursday, 12 Dec 2019, 9:30am-10:00am

Title: LSAR: Efficient Leverage Score Sampling Algorithm for the Analysis of Big Time Series Data

Abstract: We apply methods from randomized numerical linear algebra (RandNLA) to develop improved algorithms for the analysis of large-scale time series data. Among other things, we use RandNLA techniques to develop a new fast algorithm to estimate the leverage scores of an autoregressive model in big data regimes. We show that the accuracy of approximations lies within $(1 + O(\epsilon))$ of the true leverage scores with high probability. These theoretical results are exploited to develop an efficient leverage score sampling algorithm to fit an appropriate autoregressive model to big time series data and find the maximum likelihood

estimates of its parameters. We show that our pro- posed algorithm has a worst-case running time that significantly improves those of the state-of-the-art alternatives in big data regimes. Empirical results on large-scale synthetic as well as real data highly support the theoretical results and reveal the efficacy of this new approach.

11. Glen Livingston Jr, University of Newcastle, Australia

Time: Thursday, 12 Dec 2019, 10:00am-10:30am

Title: ARMA Models and Big Data

Abstract: The Autoregressive Moving Average (ARMA) model is a widely applied model for stationary time series. Within the regime of big data, estimating the parameters of an ARMA model can be costly due to the likelihood function in being a non-convex nonlinear function. To overcome this we develop an algorithm utilising the Hannan-Rissanen (HR) algorithm in conjunction with an algorithm recently developed by Eshragh et al. to analyse and fit AR models in big data regimes. The HR algorithm uses high order AR models to calculate an approximation for the time series of white noise. This is then used to estimate the MA parameters in the ARMA model. The effectiveness of this algorithm for several large-scale synthetic and real time series data will also be assessed.

12. Samudra Herath, University of Adelaide, Australia

Time: Thursday, 12 Dec 2019, 11:00am-11:30am

Title: Name-like Numbers for Simulating Names in Entity Resolution

Abstract: Accurate and efficient entity resolution (ER) has been a problem in data analysis and data mining projects for decades. It is used as a basic tool in data integration to combine multiple datasets. ER has been studied for years, but the focus has been pairwise comparisons. In our work, we are interested in the global matching problem for large datasets. Good datasets in this domain are rare, and often much smaller than needed for this type of project. Simulation is one technique to approach generating datasets for testing. However, simulation of the many individual details of identification keys is not needed when we consider the global matching problem. We need only simulate at the level of detail needed to understand distances between keys. Avoiding unnecessary detail can make the process more scalable. However, simulation need also be realistic. In this talk we will discuss how to simulate simple vectors in a space that approximates the properties of names (which are commonly used as identification keys) as one step towards being able to generate large simulated datasets for large-scale testing of global matching techniques.

13. Yang Liu, University of Queensland, Australia

Time: Thursday, 12 Dec 2019, 11:30am-12:00pm

Title: Stability Analysis of Newton-MR Under Hessian Perturbations

Abstract: Recently, stability of Newton-CG under Hessian perturbations, i.e., inexact curvature information, have been extensively studied. Such stability analysis has subsequently been leveraged in designing variants of Newton-CG in which, to reduce the computational costs involving the Hessian matrix, the curvature is suitably approximated. Here, we do that for Newton-MR, which extends Newton-CG in the same manner that MINRES extends CG. Unlike the stability analysis of Newton-CG, which relies on spectrum preserving perturbations in the sense of Lowner partial order, our work here draws from matrix perturbation theory to estimate the distance between the underlying exact and perturbed sub-spaces. Numerical experiments demonstrate great degree of stability for Newton-MR, amounting to a highly efficient algorithm in large-scale problems.

14. Russell Tsuchida, University of Queensland, Australia

Time: Thursday, 12 Dec 2019, 12:00pm-12:30pm

Title: Richer Parameter Priors for Infinitely Wide MLPs

Abstract: It is well-known that the prior distribution over functions induced through a zero-mean iid prior distribution over the parameters of a multi-layer perceptron (MLP) converges to a Gaussian process (GP), under mild conditions. We extend this correspondence firstly to the case of independent parameter priors with non-zero means, and secondly to partially exchangeable parameter priors. We discuss how the second prior arises naturally when considering an equivalence class of functions in an MLP. The model resulting from partially exchangeable priors is a GP with an additional level of inference in the sense that it requires marginalisation over hyperparameters. We evaluate the kernels of the limiting GP in deep MLPs, and show that they avoid certain pathologies present in previously studied priors. We empirically evaluate our claims by measuring the maximum mean discrepancy between finite width models and limiting models. We apply our limiting model and a GP with fixed hyperparameters to some synthetic regression problems and compare the performance of both models.

15. James Juniper, University of Newcastle, Australia

Time: Thursday, 12 Dec 2019, 2:00pm-2:30pm

Title: 'Unreasonable Effectiveness' of Machine Learning in Both the Natural Sciences and the Social Sciences?

Abstract: The paper will engage in an ontologically-based inquiry into the "unreasonable effectiveness" of machine learning within a natural science context and the "reasonable ineffectiveness" of machine learning within a social science context. To this end it will review the literature that has attempted to explain the power of deep learning recursive and convolution neural networks as approximating engines. It will also examine critical debates about kernel-based techniques and the training of parameters at the "edge of chaos". Finally, it will touch on important ontological features of relevance to the social sciences, especially those of a macroeconomic nature, that may help to explain why the current generation of machine learning models must, for the foreseeable future, face profound limitations in this particular application domain.

16. Vektor Dewanto, University of Queensland, Australia

Time: Thursday, 12 Dec 2019, 2:30pm-3:00pm

Title: A Review on Average-reward Reinforcement Learning

Abstract: This work reviews average reward methods for reinforcement learning. It covers both value- and policy-based methods, as well as both model-free and model-based methods. It also discusses some connection to works that modify the discounting in order to better approximate the average reward. We systematically evaluated the existing (major) methods on small up to large finite-MDPs. Our analysis identifies the state-of-the-art method and suggests some open problems.

17. Robert King, University of Newcastle, Australia

Time: Thursday, 12 Dec 2019, 3:00pm-3:30pm

Title: Sampling Behaviour of L-moment Estimators of the GPD Type Generalised Lambda Distribution

Abstract: The generalised lambda distribution provides a wide variety of shapes within one distributional form. The GPD type of the generalised lambda is the first type with explicit expressions for standard errors of estimates. Here we present the results of a simulation study to verify the theoretical results for these standard errors.