

2022

20-21 SEPTEMBER



ABSTRACT BOOK

**4TH IMA CONFERENCE ON THE
MATHEMATICAL CHALLENGES OF
BIG DATA**

Table of Contents

Inducement of sparsity	4
Lower dimensional representations of the filtering distribution via geometric projections	5
Shifted compression framework: generalizations and improvements	7
CLASSIX: Fast and explainable clustering based on sorting	8
High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation	9
Graph Neural Networks for Sequence Alignment	11
Error rates for kernel methods under source and capacity conditions	12
Stiffness analysis for the diffusion of fake information	13
Modularity maximisation for graphons	14
Sampling strategies for the approximation of large-scale kernel matrices	15
Asymptotic Randomised Control and Exploration Value of Learning	16
Social contagion and asset prices: Reddit's self-organised bull runs	
A dimensionality reduction technique for global optimization via random subspace embeddings	18
Constrained Optimization on Manifolds	19
Communication-efficient distributed optimization algorithms	20
High order tensor methods for non-convex unconstrained optimization	21
Generalized Diffusion for Graph Learning	23
Resolving the data ambiguity in the largest database of crystals	24
Coordinate descent on the orthogonal group for recurrent neural network training	25
Regularisation of Inverse Problems Deep Equilibrium Models versus Bi-level Optimisation	26
Lifted Bregman Networks	27
Hypergraph Generative Models and Associated Spectral Embedding Algorithms	28
A Message-Passing Approach to Epidemic Tracing and Mitigation with Apps	29
Smale's 18th problem and the barriers of deep learning	30
Normalizing Flows for Mutual Information Estimation	31
Streaming multi-resolution adaptive kernel methods	33
Neural Q-learning for solving elliptic PDEs	34
Macroprudential Policy in a Heterogeneous Environment - An Application of Agent-Based Macro Model	35

Nearest-Neighbour Resampling for Off-Policy Policy Evaluation with Applications in Market Making	36
Statistical and agent-based modelling of the transmissibility of different SARS-CoV-2 variants in England and impact of different interventions	37
Disentangling homophily, community structure and triadic closure in networks	39
Over-parametrization: Insights from solvable models	40
Computing second-order points under equality constraints using Fletcher's augmented Lagrangian	41
Higher-order signal processing with the Dirac operator	42
A subspace cubically regularised method for large-scale non-convex optimisations	43
Spectral Tools for training and Analysing for Neural Networks	44
Exploring the compositional structure of images with convolutional kernel methods	46
Accelerating Deep Unrolling Networks via Dimensionality Reduction	47
Approximation properties of two-layer neural networks with values in a Banach space	48
Identifiability in inverse reinforcement learning	49
An unsupervised manifold learning approach to untangle big data of collective human behaviors	50
Phase Collapse in Neural Networks	51
Sequential data assimilation for uncovering an influence network from Hawkes processes driven by count data	52
GNNs for Node Clustering in Signed and Directed Networks	53
Motion correction for cardiac MRI images using generative neural networks with disentangled latent space variables	54
Mirror Q-learning can exploit the structure of the environment	55
Tangent Space and Dimension Estimation with the Wasserstein Distance	56
Unifying information propagation models and their influence maximisation	58
Using Web Data and Network Science to Detect Spatial Relationships	59
Sequential Motifs in Observed Walks Through Networks	60
Robust Dynamic Network Embedding	61
The transformer leap	62

Inducement of sparsity, Heather Battey

Sparsity, the existence of many zeros or near-zeros in some domain, is widely assumed throughout the high-dimensional literature and plays at least two roles depending on context. Parameter orthogonalisation (Cox and Reid, 1987) is presented as inducement of population-level sparsity. The latter is taken as a unifying theme for the talk, in which sparsity-inducing parameterisations or data transformations are sought. Three recent examples are framed in this light: sparse parameterisations of covariance models; systematic construction of factorisable transformations for the elimination of nuisance parameters; and inference in high-dimensional regression. The solution strategy for the problem of exact or approximate sparsity inducement appears to be context specific and may entail, for instance, solving one or more partial differential equation, or specifying a parameterised path through transformation or parameterisation space.

Lower dimensional representations of the filtering distribution via geometric projections

Cohen, S. N., ¹*Fausti, E.

Mathematical Institute, University of Oxford, United Kingdom

The stochastic filtering problem consists in finding the best estimate for the true value of a system when only incomplete, noisy observations are available. In the general case, its solution is infinite-dimensional, and the filtering equations are often considered intractable from a practical perspective, which led to many approximation methods being proposed. The ‘projection filter’ is a low dimensional approximation of the optimal nonlinear filter introduced by Brigo, Hanzon and Le Gland [1]. It consists in endowing the space of probability distributions with a Riemannian metric, and using this metric to project the infinite dimensional dynamics of the nonlinear filter onto the tangent space of a suitable low dimensional submanifold.

In this paper, we focus on the filtering setting where the signal process is a continuous time homogeneous Markov chain on a finite state space, so that the optimal filter takes the simplified form of the Wonham filter. We prove a contraction result for the Wonham stochastic differential equation in the Hilbert projective metric. This allows us to prove error bounds for the projection filter, dependent on the choice of metric and of submanifold in the space of probability distributions. All results are supported by numerical experiments.

[1] Brigo D., Hanzon B., Le Gland F., 1999, *Approximate nonlinear filtering by projection on exponential manifolds of densities*, BERNOLLI, Vol: 5, Pages: 495-534, ISSN: 1350-7265

Shifted compression framework: generalizations and improvements

¹*Shulgin, E., ¹Richtárik, P.

*lead presenter,

¹King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Communication is one of the key bottlenecks in the distributed training of large-scale machine learning models, and lossy compression of exchanged information, such as stochastic gradients or models, is one of the most effective instruments to alleviate this issue. Among the most studied compression techniques is the class of unbiased compression operators with variance bounded by a multiple of the square norm of the vector we wish to compress. By design, this variance may remain high, and only diminishes if the input vector approaches zero. However, unless the model being trained is overparameterized, there is no a-priori reason for the vectors we wish to compress to approach zero during the iterations of classical methods such as distributed compressed SGD, which has adverse effects on the convergence speed. Due to this issue, several more elaborate and seemingly very different algorithms have been proposed recently, with the goal of circumventing this issue. These methods are based on the idea of compressing the *difference* between the vector we would normally wish to compress and some auxiliary vector that changes throughout the iterative process. In this work we take a step back, and develop a unified framework for studying such methods, both conceptually and theoretically. Our framework incorporates methods compressing both gradients and models, using unbiased and biased compressors, and sheds light on the construction of the auxiliary vectors. Furthermore, our general framework can lead to the improvement of several existing algorithms, and can produce new algorithms. Finally, we performed several numerical experiments to illustrate and support our theoretical findings.

CLASSIX: Fast and explainable clustering based on sorting

¹*Xinye Chen, ²Stefan Güttel

*lead presenter

¹ The University of Manchester, United Kingdom

² The University of Manchester, United Kingdom

We introduce a fast and explainable clustering method called CLASSIX. The method shares features with both distance and density based clustering methods and consists of two phases, namely a greedy aggregation phase of the sorted data into groups of nearby data points, followed by the merging of groups into clusters. The aggregation is based on the sorting of the data points along their first principal component, i.e. a sorting by scalar values irrespective of the data dimensionality d . This results in a significant performance advantage over tree-based query structures such as R* or ball trees which work on the d -dimensional data points and are nontrivial to implement efficiently. CLASSIX is controlled by essentially two parameters, namely a distance parameter for the aggregation and another parameter specifying the minimal cluster size. We analyse a simple Gaussian model to understand how the number of pairwise distance computations, and hence the overall time complexity, depends on these parameters. Through experiments on several benchmarks, we demonstrate that CLASSIX is competitive with state-of-the-art clustering algorithms. In particular, the algorithm exhibits linear space complexity and near-linear time complexity on a wide range of problems.

High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation

Jimmy Ba¹, Murat A. Erdogdu¹, Taiji Suzuki², Zhichao Wang³, Denny Wu¹, Greg Yang⁴

¹University of Toronto and Vector Institute. ²University of Tokyo and RIKEN AIP.

³University of California, San Diego. ⁴Microsoft Research.

We consider the following fully-connected two-layer neural network (NN) with N neurons,

$$f_{\text{NN}}(\mathbf{x}) = \frac{1}{\sqrt{N}} \sum_{i=1}^N a_i \sigma(\langle \mathbf{x}, \mathbf{w}_i \rangle) = \frac{1}{\sqrt{N}} \mathbf{a}^\top \sigma(\mathbf{W}^\top \mathbf{x}),$$

where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{W} \in \mathbb{R}^{d \times N}$, $\mathbf{a} \in \mathbb{R}^N$, σ is a nonlinearity applied entry-wise, and the training objective is to minimize the empirical risk computed on n training points. Our analysis will be made in the *proportional asymptotic limit*, i.e., the number of training data n , the input dimensionality d , and the number of features (neurons) N jointly tend to infinity. In this regime, the performance of *random features* (RF) kernel regression (where $\mathbf{W}$ is randomly initialized and only $\mathbf{a}$ is learned) has been precisely characterized [LLC18, MM19].

While these asymptotic analyses reveal interesting phenomena also present in practical deep learning, RF and kernel models do not fully explain the empirical success of neural networks: one crucial advantage of deep learning is the *ability to learn useful features* that “adapt” to the learning problem [Suz18, GMMM19]. In our setting this adaptivity arises from the learning of first-layer weights $\mathbf{W}$, which can make a difference even in the “early phase” of gradient descent training. Our goal is to rigorously characterize the presence of feature learning in the proportional limit, and demonstrate its improvement over the initial (fixed) kernel.

Our Results. We investigate arguably the most simplified scenario of feature learning: how *the first gradient step* on the first-layer parameters $\mathbf{W}$ impacts the representation of the two-layer NN. Specifically, we consider the regression setting with the MSE loss, and a student-teacher model in the proportional asymptotic limit; we precisely characterize the prediction risk of the kernel ridge regression estimator on top of the first-layer CK feature $\mathbf{x} \rightarrow \sigma(\mathbf{W}^\top \mathbf{x})$, before and after the gradient step with learning rate η .

When the target function (teacher) f^* is a single-index model, it is known that the prediction risk of a large class of RF/kernel ridge regression estimators is lower-bounded by the L^2 -norm of the “nonlinear” component the teacher $\|\mathbf{P}_{>1} f^*\|_{L^2}^2$, i.e., they can only learn *linear* functions on the input [BMR21].

one gradient step on $\mathbf{W}$, we compute the CK ridge estimator using separate training data, and compare its prediction risk against this linear lower bound. Our analysis will be made under the following two choices of learning rate scaling (see Figure 1):

- **Small lr:** $\eta = \Theta(1)$. We extend the *Gaussian Equivalence Theorem* (GET) in [HL20] to the feature map trained via multiple gradient descent steps on $\mathbf{W}$; this allows us to precisely characterize the prediction risk using random matrix theoretical tools. We prove that after one gradient step, the ridge regression estimator on the learned CK features already exhibits nontrivial improvement over the initial RF ridge model, but it remains in the “linear regime” and cannot outperform the best linear estimator on the input.
- **Large lr:** $\eta = \Theta(\sqrt{N})$. As for larger learning rate that coincides with the *maximal update parameterization* in [YH20], we prove that for certain target functions f^* , the kernel ridge regression estimator after one feature learning step can achieve lower risk than the linear lower bound $\|\mathbf{P}_{>1} f^*\|_{L^2}^2$, and thus outperform a wide range of RF and kernel models.

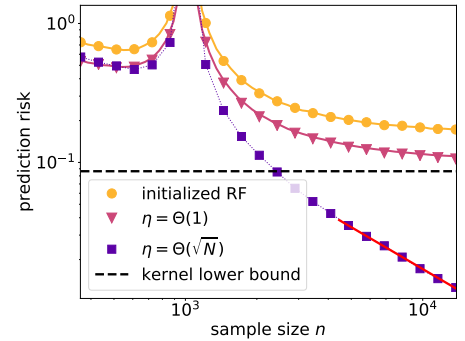


Figure 1: Prediction risk of ridge regression on trained features (erf) after one gradient step. Dots represent empirical simulations and solid lines are predicted asymptotic values; red line indicates $\Theta(d/n)$ rate.

References

- [BMR21] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin, *Deep learning: a statistical viewpoint*, arXiv preprint arXiv:2103.09177 (2021).
- [GMMM19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari, *Limitations of lazy training of two-layers neural networks*, arXiv preprint arXiv:1906.08899 (2019).
- [HL20] Hong Hu and Yue M Lu, *Universality laws for high-dimensional learning with random features*, arXiv preprint arXiv:2009.07669 (2020).
- [LLC18] Cosme Louart, Zhenyu Liao, and Romain Couillet, *A random matrix approach to neural networks*, The Annals of Applied Probability **28** (2018), no. 2, 1190–1248.
- [MM19] Song Mei and Andrea Montanari, *The generalization error of random features regression: Precise asymptotics and double descent curve*, arXiv preprint arXiv:1908.05355 (2019).
- [Suz18] Taiji Suzuki, *Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality*, arXiv preprint arXiv:1810.08033 (2018).
- [YH20] Greg Yang and Edward J Hu, *Feature learning in infinite-width neural networks*, arXiv preprint arXiv:2011.14522 (2020).

Graph Neural Networks for Sequence Alignment

Christopher PolICASTRO

University of Oxford

Sequential data occurs in different areas of academia, industry and government. It can occur in a categorical format like text in humanities or nucleotides in biology. It can occur in a numerical format like stock prices in finance or temperature in climatology. Regardless of the format, sequential data tends to divide into components like a trend component, a periodic component and a residual component.

The periodic component contains recurring patterns. For example, temperature in climatology exhibits seasonal variations repeating year over year. The patterns can recur within a dataset or between datasets as motifs. Determination of motifs in the periodic component can help with descriptive tasks and predictive tasks in the processing of sequential data.

However, a parametric model can fail to identify motifs in corrupted data. Large, diverse and intricate datasets can contain anomalous values, missing values, change-points, irregular sampling, or high dimension features.

We propose a non-parametric model through the conversion of the sequence to a network. The nodes of the network are values in the sequence. The edges in the network reflect the relationship between values through a magnitude and phase.

Conversion of the sequence to network allows us to leverage network analysis for determination of motifs. While we could use spectral methods to capture information from the magnitude through the Laplacian, we can capture information from the magnitude and phase pivoting from the Laplacian to the Magnetic Laplacian developed by Fanuël and collaborators.

Since the spectral methods disallow the covariates like the values in the sequence, we propose a variant of the MagNet graph neural network developed by He and collaborators. The architecture allows us to implement node classification for determination of motifs. Experiments on real and synthetic datasets show the utility of the approach.

Title: Error rates for kernel methods under source and capacity conditions

Abstract:

We investigate the rates of decay of the prediction error for kernel methods under the Gaussian design and source/capacity assumptions. For kernel ridge regression, we derive all the observable rates, and characterize the regimes in which each hold. In particular, we show that the decay rate may transition from a fast, noiseless rate to a slow, noisy rate as the sample complexity is increased. For noiseless kernel classification, we derive the rates for two standard classifiers, margin-maximizing SVMs and ridge classifiers, and contrast the two methods. In both cases, the derived rates also describe to a good degree the learning curves of a number of real datasets. This is joint work with Bruno Loureiro, Florent Krzakala and Lenka Zeborová.

Stiffness analysis for the diffusion of fake information

¹Conte, D., ²D'Ambrosio, R., ^{1*}Giordano, G., & ¹Paternoster, B.

¹[University of Salerno](#), Italy

²University of L'Aquila, Italy

The continuous development of the web and the consequent birth of social media have allowed an easier transmission of information. Therefore, nowadays users consider social media as their primary source of information. Clearly, consulting news on the web can be considered a double-edged sword, due to the propagation of fake information. Fake news can be defined as counterfeit information, created by completely ignoring the editorial rules and processes adopted in the media to ensure their verifiability.

Our aim is to model the spread of fake news through the stiffness ratio of the classical SIR model, both deterministic and stochastic. In particular, our analysis will show that, the more the stiffness index of the differential problem is larger, the faster is the transit of fake news in the exposed country. Conversely, low stiffness values correspond to a slower transit of fake news in the country. Numerical experiments, performed on real data, support the effectiveness of the approach

References:

- [1] Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. Proceedings of the royal society of London. Series A, Containing papers of a mathematical and physical character, 115(772), 700-721.
- [2] D'Ambrosio, R., Giordano, G., Mottola, S., & Paternoster, B. (2021). Stiffness analysis to predict the spread out of fake information. Future Internet, 13(9), 222.

Modularity maximisation for graphons

¹*Florian Klimm, ²Nick S. Jones & ³Michael T. Schaub

*lead presenter

¹ Max-Planck Institute for Molecular Genetics & Freie University Berlin, Germany

²Imperial College London, United Kingdom

³RWTH Aachen University, Germany

Networks are a widely-used tool to investigate the large-scale connectivity structure in complex systems and *graphons* have been proposed as an infinite size limit of dense networks. The detection of communities or other meso-scale structures is a prominent topic in network science as it allows the identification of functional building blocks in complex systems. When such building blocks may be present in graphons is an open question. We define a graphon-modularity and demonstrate that it can be maximised to detect communities in graphons. We then investigate specific synthetic graphons and show that they may show a wide range of different community structures. We also reformulate the graphon-modularity maximisation as a continuous optimisation problem and so prove the optimal community structure or lack thereof for some graphons, something that is usually not possible for networks. Furthermore, we demonstrate that estimating a graphon from network data as an intermediate step can improve the detection of communities, in comparison with exclusively maximising the modularity of the network. While the choice of graphon-estimator may strongly influence the accord between the community structure of a network and its estimated graphon, we find that there is a substantial overlap if an appropriate estimator is used. Our study demonstrates that community detection for graphons is possible and may serve as a privacy-preserving way to cluster network data.

See preprint [arXiv:2101.00503](https://arxiv.org/abs/2101.00503) for details.

Sampling strategies for the approximation of large-scale kernel matrices

¹*Bertrand Gauthier

*lead presenter

¹ Cardiff University, United Kingdom

This title is intended for an in-person oral presentation.

Integral operators with positive-semidefinite (PSD) kernels play a central role in kernel methods, a class of learning and modelling techniques exploiting the properties of reproducing kernel Hilbert spaces (RKHSs); this class of operators encompasses the PSD matrices, emphasising its widespreadness in Mathematics and Data Science. Sparse and low-rank approximation schemes for this type of operators and matrices are essential for the implementation, acceleration and scaling-up of numerical strategies involving PSD kernels or large-scale PSD matrices.

In this talk, we will give an overview of the connections between the approximation of integral operators with PSD kernels, the low-rank approximation of positive-semidefinite matrices, and the kernel embedding of measures. We will in particular describe how trace-class integral operators defined from a PSD kernel K can be isometrically represented as potentials in the RKHS associated with the squared kernel $|K|^2$. We will then present various sampling-based approximation strategies that can be implemented from the considered isometric representation.

This presentation is based on the following preprints:

1. B. Gauthier. *Nyström approximation and reproducing kernels: embeddings, projections and squared-kernel discrepancy*. <https://hal.archives-ouvertes.fr/hal-03207443/>
2. M. Hutchings & B. Gauthier. *Local optimisation of Nyström samples through stochastic gradient descent*. <https://arxiv.org/abs/2203.13284>

Asymptotic Randomised Control and Exploration Value of Learning

Tanut Treetanthiploet¹

The Alan Turing Institute

The Optimistic Principle (or the Upper-Confidence-Bound) is one of the classical approach where the decision maker often uses to explore the unknown environment. The core idea of this principle is to choose the option in favour of uncertainty to help them learn which contrast to most economic theory where human often prefers less uncertain choice. This raises a natural question what the exploration value for learning should correspond to?

In this talk, we will consider the correlated multi-armed bandit as a regularised control problem using Bayesian inference. We will show through the asymptotic analysis that the (near-)optimal solution can be described as a semi-index strategy whose value can be decomposed as the summation between our expected gain and the improvement of the value function describing the exploration-exploitation trade-off. Our exploration value reflects the asymmetry of available information from each option and the curvature of our reward. Our result is closely related to the optimistic principle. However, instead of making decision in favour of uncertainty, our result suggests to make decision in favour of the reduction of uncertainty.

This talk based on the joint work with Samuel Cohen.

Social contagion and asset prices: Reddit's self-organised bull runs

Valentina Semenova^{* 1,3} and Julian Winkler^{2,3}

¹*Department of Mathematics, University of Oxford -*

²*Department of Economics, University of Oxford*

³*Institute of New Economic Thinking at the Oxford Martin School*

April 28, 2022

Abstract

This paper develops an empirical and theoretical case for how ‘hype’ among retail investors can drive large asset price fluctuations. We use text data from discussions on WallStreetBets (WSB), an online investor forum with over eleven million followers as of February 2022, as a case study to demonstrate how retail investors influence each other, and how social behaviours impact financial markets. We document that WSB users adopt price predictions about assets (bullish or bearish) in part due to the sentiments expressed by their peers. Peer influence is estimated using random, temporal variation in peers and an interaction network approach for two different identification strategies. Discussions about stocks are also self-perpetuating: narratives about specific assets spread at an increasing rate before peaking, and eventually diminishing in importance – a pattern reminiscent of an epidemiological setting. To consolidate these findings, we develop a model for the impact of social dynamics among retail investors on asset prices. We find that the interplay between ‘trend following’ and ‘consensus formation’ determines the stability of price returns, with socially-driven investing potentially causing oscillations and cycles. Our framework helps identify components of asset demand stemming from social dynamics, which we predict using WSB data. Our predictions explain significant variation in stock market activity. These findings emphasise the role that social dynamics play in financial markets, amplified by online social media.

Speaker: Adilet Otemissov

Mode: Remote (Online) presentation

Joint with: Coralia Cartis and Estelle Massart

Title: A dimensionality reduction technique for global optimization via random subspace embeddings

Abstract:

In an attempt to improve the scalability of global optimization problems, we propose a general random subspace algorithmic framework, which tackles the (high-dimensional) global optimization problem by repeatedly and possibly adaptively solving reduced subproblems with the domains restricted to be along randomly embedded low-dimensional subspaces. We analyse the convergence of the proposed framework for Lipschitz continuous functions using tools from conic integral geometry and random matrix theory. We then particularise the framework and analysis for the class of functions with low effective dimensionality, which are constant along an (unknown) linear subspace and only vary over the effective (complement) subspace. We show that for these functions the convergence could potentially be exponentially better and that, under certain assumptions, it does not depend on the ambient dimension. Our numerical experiments on functions with low effective dimensionality illustrate two things: 1) the improved scalability of the framework when coupled with state-of-the-art global — and even local — optimization solvers for the subproblems 2) the ability of the framework to effectively estimate the (initially unknown) effective dimensionality of the objective function.

Further references:

<https://arxiv.org/abs/2003.09673>

<https://arxiv.org/abs/2009.10446>

<https://arxiv.org/abs/2107.12102>

Constrained Optimization on Manifolds

Melanie Weber

Oxford, Mathematical Institute

Many applications involve non-Euclidean data, such as graphs, strings, or matrices. In such cases, exploiting Riemannian geometry can deliver algorithms that are computationally superior to standard nonlinear programming approaches. This observation has resulted in an increasing interest in Riemannian methods in the optimization and machine learning community. In this talk, we consider the problem of optimizing a convex function on a (Riemannian) manifold subject to convex constraints. Several classical problems that arise in Machine Learning applications can be phrased as constrained optimization tasks on manifolds. This includes matrix-valued tasks, such as barycenter problems, as well as the computation of Brascamp-Lieb constants. The latter is of central importance in many areas of mathematics and computer science through connections to maximum likelihood estimators in Gaussian models, Tyler's M-estimator of scatter matrices or the computation of marginals in quantum information theory. In this talk, we present algorithms for solving constrained optimization problems on manifolds and discuss applications to the two problems described above. Based on joint work with Suvrit Sra.

Communication-efficient distributed optimization algorithms

¹*Condat, L. & ¹Richtárik, P.

*lead presenter. Contact: see <https://lcondat.github.io/>

¹King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

In the big data era, the explosion in size and complexity of the data arises in parallel to a shift towards distributed computations, as modern hardware increasingly relies on the power of uniting many parallel units into one system. For distributed optimization and learning tasks, specific issues arise, such as decentralized data storage. In the modern paradigm of federated learning, a potentially huge number of devices, with their owners' data stored on each of them, are involved in the collaborative process of training a global machine learning model. But users are increasingly sensitive to privacy concerns and prefer their data to never leave their devices. Thus, the devices have to **communicate** the right amount of information back and forth with a distant server, for this distributed learning process to work and converge to a global solution.

Communication, which can be costly and slow, is the main bottleneck in this framework. So, it is of primary importance to devise novel algorithmic strategies, with reduced computation and communication complexities. A first strategy is to communicate less frequently than at every iteration, making use of so-called local steps. A second strategy is to compress the communicated vectors.

In this talk, I will present several recent contributions we did in our *Optimization and Machine Learning* group in this realm of communication-efficient distributed optimization algorithms, with proved convergence guarantees.

High order tensor methods for non-convex unconstrained optimization

*Kate Wenqi Zhu, [†]Coralia Cartis

Some recent works, such as [1, 2, 3, 4], have shown that some optimization algorithms have better worst-case evaluation complexity when using higher-order derivative information of the objective function, beyond second order derivatives, together with regularization techniques. We give a brief description of the methods as follows. Given $p \geq 2$, we consider the optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{0.1}$$

where f is p times continuously differentiable. Let $T_p(x, s)$ be the Taylor series of the function $f(x + s)$ at x truncated at order p ,

$$T_p(x, s) := f(x) + \sum_{j=1}^p \frac{1}{j!} \nabla_x f_x^j(x) [s]^j, \tag{0.2}$$

where the notation $T[s]^j$ denotes the tensor T applied j times to the vector s . We aim to approximately minimize the regularized Taylor series model over s

$$m(x, s, \sigma) = T_p(x, s) + \frac{\sigma}{p+1} \|s\|^{p+1} \tag{0.3}$$

References

- [1] Birgin, E. G., et al. “Worst-Case Evaluation Complexity for Unconstrained Nonlinear Optimization Using High-Order Regularized Models.” *Mathematical Programming*, vol. 163, no. 1-2, 2016, pp. 359–368., <https://doi.org/10.1007/s10107-016-1065-8>.
- [2] Cartis, C., et al. “A Concise Second-Order Complexity Analysis for Unconstrained Optimization Using High-Order Regularized Models.” *Optimization Methods and Software*, vol. 35, no. 2, 2019, pp. 243–256., <https://doi.org/10.1080/10556788.2019.1678033>.
- [3] Cartis, Coralia, et al. “Sharp Worst-Case Evaluation Complexity Bounds for Arbitrary-Order Nonconvex Optimization with Inexpensive Constraints.” *SIAM Journal on Optimization*, vol. 30, no. 1, 2020, pp. 513–541., <https://doi.org/10.1137/17m1144854>.
- [4] Nesterov, Yurii. “Implementable Tensor Methods in Unconstrained Convex Optimization.” *Mathematical Programming*, vol. 186, no. 1-2, 2019, pp. 157–183., <https://doi.org/10.1007/s10107-019-01449-1>.

Generalized Diffusion for Graph Learning

¹Csaba Toth, ^{*}^{2,3}Darrick Lee, ²Celia Hacker, & ¹Harald Oberhauser^{*}lead presenter,

¹ University of Oxford, UK

² École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

³ University of Oxford, UK (beginning August 1, 2022)

Obtaining an expressive representation of labelled graphs has become increasingly important in a variety of application domains such as biology, chemistry, and social/knowledge networks. One way to construct such a representation is to first build node features by iteratively aggregating information about local neighbourhood structures, and then using the collection of such node features to get a latent description of the whole graph. Random walks and graph diffusion are two methods to summarize local information, which have recently been highly successful in a variety of graph learning tasks.

In this work, we introduce a generalized diffusion model which combines the random walk and classical diffusion perspective on graph learning. We establish the mathematical framework behind this generalized diffusion model and develop efficient algorithms to compute the resulting representation. Finally, we include an ablation study that demonstrates how this approach extends and improves upon graph diffusions, as well as benchmark it against other recent methods.

Resolving the data ambiguity in the largest database of crystals

Kurlin, V.A. [_____](#)

Materials Innovation Factory, University of Liverpool, UK

A periodic crystal can be represented by many minimal cells of atoms or molecules that are periodically repeated in three directions. Since crystal structures are determined in a rigid form, their strongest equivalence is rigid motion or isometry also including reflections. Their classification up to isometry needs strong invariants [1] and metrics that are continuous under perturbations because any conventional cell may double in size under small perturbations [2].

The recent Pointwise Distance Distributions are ultra-fast isometry invariants of periodic point sets. More than 200 billion pairwise comparisons between all 660K+ periodic crystals (full 3D structure; no disorder) in the Cambridge Structural Database (CSD) were completed over two days on a modest desktop computer. These computations detected five pairs with identical geometry but different chemistry [3, sec. 7]. Since these coincidences are physically impossible, five journals are now investigating the integrity of the underlying publications.

The above experiments justified the Crystal Isometry Principle, which states that any periodic crystal is uniquely determined by its geometric structure. Replacing any atom with a different one inevitably perturbs distances to neighbouring atoms, which is detected by our invariants.

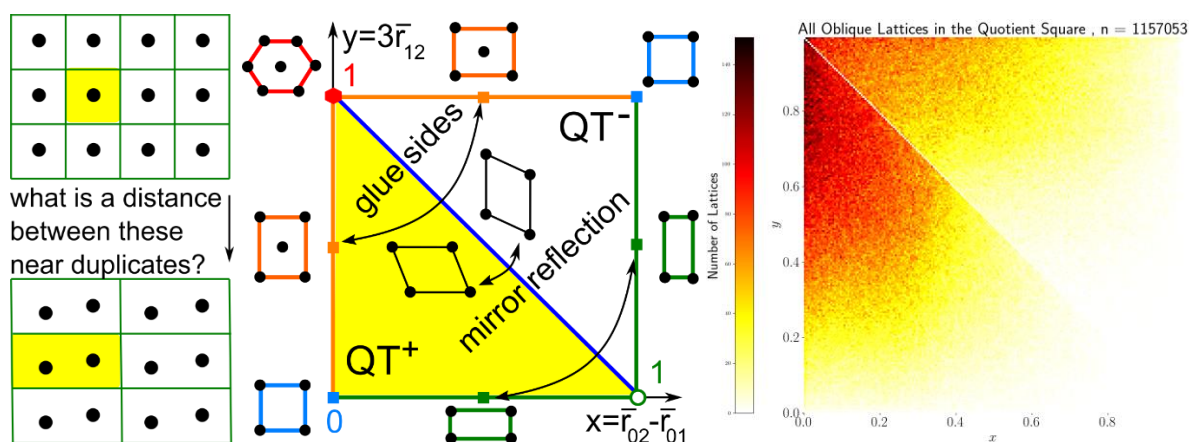


Figure 1. Left: discontinuity of cells under tiny perturbations. Middle: 2D lattice space up to isometry and scaling [4]. Right: density map of oblique 2D lattices in CSD periodic crystals.

Then all crystals live in a common Crystal Isometry Space (CRISP), so Mendeleev's periodic table representing individual elements, categorised by atomic number and group, extends into a continuous space for all solid crystalline materials. Diamond and graphite, both consisting purely of carbon, have different locations in CRISP, see all CSD lattice projections in Fig. 1.

[1] P.Sacchi et al. Same or different – that is the question: identification of crystal forms from crystal structure data. CrystEngComm 22 (43), 7170-7185, 2020. [2] A.Pulido et al. Functional materials discovery using energy-structure maps. Nature 543, 657-664, 2017.

[3] D.Widdowson et al. Average Minimum Distances – fundamental invariants for mapping all periodic crystals. MATCH Comm. in Math. Comp. Chemistry, v.87(3), p.529-559, 2022.

[4] Kurlin, V. Mathematics of 2-dimensional lattices. <https://arxiv.org/abs/2201.05150>.

Coordinate descent on the orthogonal group for recurrent neural network training

¹*Massart, E., ²Abrol, V.

*Lead presenter

¹FNRS scientific collaborator, UCLouvain, Belgium

²Infosys Centre for AI, IIT Delhi, India

Recurrent neural networks (RNNs) are often subject to vanishing or exploding gradients. These two issues can be alleviated by ensuring orthogonality of the recurrent matrix, resulting in an additional computational burden when training the model. This motivated several recent works to propose parametrizations of the set of orthogonal matrices that lead to efficient and easy-to-implement optimization algorithms. Most of the existing methods suffer either from a lack of theoretical support, or a lack of practicality (with iteration costs increasing cubically with the number of hidden units).

We propose to use a stochastic Riemannian coordinate descent algorithm on the orthogonal group for RNN training. Coordinate descent amounts to rotate successively two columns of the recurrent matrix, an operation whose cost is linear in the size of the matrix. In the case when the coordinate is selected uniformly at random at each iteration, we prove the convergence of the proposed algorithm under standard assumptions on the loss function, stepsize and minibatch noise. In addition, we numerically demonstrate that the Riemannian gradient in recurrent neural network training has an approximately sparse structure. Leveraging this observation, we propose a faster variant of the proposed algorithm that relies on the Gauss-Southwell rule. Experiments on a benchmark recurrent neural network training problem are presented to demonstrate the effectiveness of the proposed algorithm.

This abstract relies on our recent work “E. Massart, V. Abrol, Coordinate descent on the orthogonal group for recurrent neural network training, AAAI 2022”.

Title

Regularisation of Inverse Problems: Deep Equilibrium Models versus Bi-level Optimisation

Abstract

Variational regularisation methods are commonly used to approximate solutions of inverse problems. In recent years, model-based variational regularisation methods have often been replaced with data-driven ones such as the fields-of-expert model (Roth and Black, 2009). Training the parameters of such data-driven methods can be formulated as a bi-level optimisation problem. In this talk, we compare the framework of bi-level optimisation for the training of data-driven variational regularisation models with the novel framework of deep equilibrium models (Bai, Kolter, and Koltun, 2019) that has recently been introduced in the context of inverse problems (Gilton, Ongie, and Willett, 2021). We show that computing the lower-level optimisation problem within the bi-level formulation with a fixed-point iteration is a special case of the deep equilibrium framework. We compare both approaches computationally, with a variety of numerical examples for the inverse problems of denoising, inpainting and deconvolution.

Authors

Danilo Riccio, Queen Mary University of London, United Kingdom, _____

Matthias J. Ehrhardt, University of Bath, United Kingdom, [Martin Benning](#), Queen Mary University of London, United Kingdom,

Lifted Bregman Networks

^{1*}Wang, X., ²Benning, M.

*lead presenter

¹, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK

² School of Mathematical Sciences, Queen Mary University of London, UK

A combination of gradient-based optimisation method and gradient computation via backpropagation is often the method of choice when it comes to training deep neural networks. In this work, we focus on alternative training strategies that are based on recent developments in distributed optimisation of deeply nested systems (Carreira-Perpinan & Wang 2014). In particular, we propose a novel training approach that is based on proximal activation functions and generalised Bregman distances. The main advantages of this approach compared to previous approaches is that it enables (potentially distributed) training of neural network parameters without having to differentiate activation functions.

Hypergraph Generative Models and Associated Spectral Embedding Algorithms

Xue Gong^{1,2*}, Desmond John Higham¹, and Konstantinos Zygalakis¹

¹ School of Mathematics, University of Edinburgh, Edinburgh, EH9 3FD, UK

² The Maxwell Institute for Mathematical Sciences, EH8 9BT, UK

Graphs are ubiquitous in many real-world scenarios, where nodes represent agents and edges capture interactions between pairs of agents. When the interactions involve groups of agents, for example, meetings between more than two people, the pairwise relationship is insufficient to reflect some real-world phenomena. Hypergraphs capture these higher-order interactions through hyperedges that may link more than two nodes.

We consider the problem of mapping nodes in a hypergraph into low-dimensional Euclidean space such that most interactions are short-range. This embedding is relevant to many follow up tasks, such as node reordering, node clustering, and visualization. We focus on hypergraph embedding algorithms to recover linear and periodic structures. In the periodic case, we allow for wrap-around in the distance measure. We show that two spectral hypergraph embedding algorithms, which reveal linear and periodic structures respectively, are associated with a new class of hypergraph generative models. The models generate each hyperedges according to node positions in the embedded space and encourage short-range connections. This pattern is created by assigning each hyperedge a probability that decays with the sum of the squared pairwise distances between nodes in the hyperedge. This random graph model allows us to quantify the relative presence of periodic and linear structures in the data through maximum likelihood. We demonstrate this approach on synthetic and real-world hypergraphs.

Acknowledgements X.G. acknowledges support of MAC-MIGS CDT under EPSRC grant EP/S023291/1. D.J.H. was supported by EPSRC Programme Grant EP/P020720/1.

A Message-Passing Approach to Epidemic Tracing and Mitigation with Apps

^{1,2} Ginestra Bianconi, ^{*2} Hanlin Sun, ³ Giacomo Rapisardi, ⁴ Alex Arenas

¹The Alan Turing Institute, United Kingdom

² Queen Mary University of London, United Kingdom. [_____](#)

³ Barcelona Supercomputing Center (BSC) and Universitat Rovira i Virgili, Spain.

⁴ Universitat Rovira i Virgili, Spain

With the hit of new pandemic threats, scientific frameworks are needed to understand the unfolding of the epidemic. The use of mobile apps that are able to trace contacts is of utmost importance in order to control new infected cases and contain further propagation. Here we present a theoretical approach [1] that uses both percolation and message-passing techniques to quantify the role of automated contact tracing in mitigating an epidemic wave.

Our approach captures the steady state of the SIR epidemic spreading with contact-tracing based on the mapping of the process to link percolation. Assuming perfect efficiency of the app, the model is based on the fact that only individuals with the app infected by individuals with the app are not able to spread the disease further if infected (see Figure 1). Our study goes beyond previous attempts to study the effect of the app by fully capturing the non-linear effects of the dynamics and the phase diagram of the process. Moreover, we show that the adoption of the app by a large fraction of the population increases the value of epidemic threshold, and the best strategy to mitigate the epidemic spreading is targeting the hubs. The analytical results are compared with extensive Monte Carlo simulations showing good agreement for homogeneous, heterogeneous networks models and for real data.

In conclusion, the proposed theoretical framework is able to assess the expected impact of contact-tracing apps in the course of an epidemic capturing the non-linear effect of the spreading dynamics. These results are important to quantify the level of adoption needed for contact-tracing apps to be effective in mitigating an epidemic.

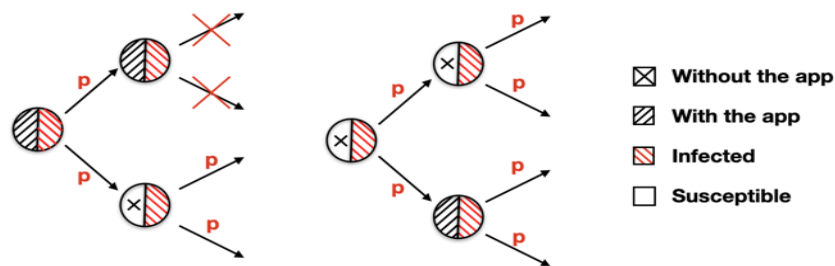


Figure 1: Sketch of the infection pathways that leads to the epidemic spreading in a population in which there are individuals that have and have not adopted the app.

Reference

[1] Bianconi, G., Sun, H., Rapisardi, G., & Arenas, A. (2021). Message-passing approach to epidemic tracing and mitigation with apps. *Physical Review Research*, 3(1), L012014.

Smale's 18th problem and the barriers of deep learning

^{1*}Colbrook, M.J., ²Antun, V., ³Hansen, A.C.

^{*}lead presenter

¹École Normale Supérieure, France²University of Oslo, Norway

³University of Cambridge, UK

Deep learning (DL) has had unprecedented success and is now rapidly entering scientific computing (SC). However, DL suffers from a universal phenomenon: instability, despite universal approximation results that often guarantee the existence of stable and accurate neural networks (NNs). We show the following paradox [1,2]. There are well-conditioned problems in SC where one can prove the existence of NNs with great approximation qualities, however, there does not exist any algorithm that can train such a NN. For any positive integers $K > 2$ and L , there are cases where simultaneously: (a) no algorithm can train a NN correct to K digits, (b) there exists an algorithm that trains a NN with $K - 1$ correct digits, but any such algorithm needs arbitrarily many training data, (c) there exists an algorithm that trains a NN with $K - 2$ correct digits using no more than L training samples. These results provide basic foundations for Smale's 18th problem and imply a classification theory describing conditions under which (stable) NNs with a given accuracy can be trained. We begin this theory by initiating a unified theory for compressed sensing and DL, leading to sufficient conditions for the existence of training algorithms for stable NNs in inverse problems. We introduce FIRENETs, which we prove and numerically verify are stable. FIRENETs only require $O(\log(1/\varepsilon))$ hidden layers for an ε -accurate solution to the inverse problem.

[1] M.J. Colbrook, V. Antun, and A.C. Hansen, "The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem," PNAS 119.12 (2022).

[2] C.Q. Choi, "Some AI Systems May Be Impossible to Compute," IEEE Spectrum, <https://spectrum.ieee.org/deep-neural-network>

Normalizing Flows for Mutual Information Estimation

Ni, Haoran and Lotz, Martin

Warwick University

Mutual information, as a measure of mutual dependence of random quantities without specific modelling assumptions, plays an important role in many scientific disciplines. However, estimating mutual information numerically from high-dimensional data remains a difficult problem. Classic non-parametric estimators of entropy and mutual information depend either on density estimation and Monte Carlo integration, or on the computation of k nearest neighbours (kNN). The first type of algorithm suffers from the curse of dimensionality, while kNN methods lack accuracy in the presence of correlation. Non-parametric estimators are usually fast and accurate without any further assumptions, but they do not scale well for small-sized samples in high-dimensional spaces. This impediment has motivated the parametric estimators, which take advantage of the expressive power of deep learning architectures in parameterizing the space of functions, to find the variational bounds for mutual information. Some recent research include MINE (Belghazi et al., 2018; Song and Ermon, 2019), Contrastive Predictive Coding (van den Oord et al., 2018), DEMI (Liao et al., 2020) and Normalizing Flows based entropy estimators (Ao and Li, 2021). As shown by Song and Ermon (2019), these models based on discriminative learning can introduce high variance in the estimation, while “generative” models such entropy estimators based on normalizing flows need to learn three distributions to compute mutual information. We propose a principled mutual information estimator based on a generalization of normalizing flows, a deep generative model that has gained considerable popularity in recent years. This estimator not only estimates the marginal and joint entropy in a systematic way, but also uses a generative-discriminative combined structure to obtain better performance in estimating mutual information. Empirical results demonstrate that our proposed estimator exhibits improved bias-variance trade-offs on standard benchmark tasks.

References

Ziqiao Ao and Jinglai Li. Entropy estimation via normalizing flow. In *Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2022)*, Proceedings

of the AAAI Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence, November 2021. Not yet published as of 22/03/2022.; 36th AAAI Conference on Artificial Intelligence, AAAI-22 ; Conference date: 22-02-2022 Through 01-03-2022.

Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R. Devon Hjelm, and Aaron C. Courville. MINE: mutual information neural estimation. *CoRR*, abs/1801.04062, 2018. URL <http://arxiv.org/abs/1801.04062>.

Ruizhi Liao, Daniel Moyer, Polina Golland, and William M. Wells III. DEMI: discriminative estimator of mutual information. *CoRR*, abs/2010.01766, 2020. URL <https://arxiv.org/abs/2010.01766>.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *CoRR*, abs/1910.06222, 2019. URL <http://arxiv.org/abs/1910.06222>.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL <http://arxiv.org/abs/1807.03748>.

Streaming multi-resolution adaptive kernel methods

Andreas Oslandsbotn

Kernel ridge regression (KRR) is a popular kernel method for non-linear non-parametric learning. Yet, existing implementations of KRR require that all data is stored in-memory, which severely limits the use of KRR in contexts where data size far exceeds in-memory size. One approach to address this issue is the streaming model of computation.

Furthermore, in high dimensions, KRR suffers from the "curse of dimensionality" requiring large amounts of training data to converge. Moreover, the computational complexity, memory requirement, and the number of parameters to learn grow unbounded with the number of training samples, a drawback known as the "curse of kernelization". In the context of streaming, prospects of unbounded data streams makes these shortcomings even more detrimental.

In this talk we present StreaMRAK - a streaming multi-resolution version of KRR. StreaMRAK uses a streaming computational model to reduce the in-memory requirements. Furthermore, a new sub-sampling scheme is proposed to address the curses of dimensionality and kernelization. The sub-sampling is adapted to a multi-resolution scheme, which once combined with a localized kernel, gives frequency and location-based discretization, similar to the wavelet framework that has shown great success in numerous applications.

The presented numerical studies show that the proposed algorithm is fast and accurate.

Neural Q-learning for solving elliptic PDEs

Samuel N Cohen, *Deqing Jiang, Justin Sirignano

[University of Oxford, UK](#)

University of Oxford & Alan Turing Institute, UK

Solving high-dimensional partial differential equations (PDEs) is a major challenge in scientific computing. We develop a new numerical method for solving elliptic-type PDEs by adapting the Q-learning algorithm in reinforcement learning. Our “Q-PDE” algorithm is mesh-free and therefore has the potential to overcome the curse of dimensionality. Using a neural tangent kernel (NTK) approach, we prove that the neural network approximator for the PDE solution, trained with the Q-PDE algorithm, converges to the trajectory of an infinite-dimensional ordinary differential equation (ODE) as the number of hidden units goes to infinity.

For monotone PDE (i.e. those given by monotone operators, which may be nonlinear), despite the lack of a spectral gap in the NTK, we then prove that the limit neural network, which satisfies the infinite-dimensional ODE, converges in L^2 to the PDE solution as the training time goes to infinity. More generally, we can prove that any fixed point of the wide-network limit for the Q-PDE algorithm is a solution of the PDE (not necessarily under the monotone condition). The numerical performance of the Q-PDE algorithm is studied for several elliptic PDEs.

Macroprudential Policy in a Heterogeneous Environment - An Application of Agent-Based Macro Model

¹*Kaszowska-Mojša, J., ²Pipień, M.

*lead presenter¹, University of Oxford, United Kingdom² Cracow University of Economics, Poland

Assessment of welfare effects of macroprudential policy seems the most important application of the Dynamic Stochastic General Equilibrium (DSGE) framework of macro-modelling. In particular, the DSGE-3D model, with three layers of default (3D), was developed and used by the European Systemic Risk Board and European Central Bank as a reference tool to formally model the financial cycle as well as to analyze effects of macroprudential policies.

Despite the extreme importance of incorporating financial constraints in Real Business Cycle (RBC) models, the resulting DSGE-3D construct still embraces the representative agent idea, making serious analyses of diversity of economic entities impossible. We present an alternative to DSGE modelling that seriously departs from the assumption of the representativeness of agents. Within an Agent Based Modelling (ABM) framework, we build an environment suitable for performing counterfactual simulations of the impact of macroprudential policy on the economy, financial system and society.

We contribute to the existing literature by presenting an ABM model with broad insight into heterogeneity of agents. Including the heterogeneity of agents in the ABM macro model requires the analyses of large data sets to approximate multivariate distributions and the use of big data techniques. In our paper, we focus on both the methodological aspects (calibration, validation and testing) and the applications. In particular, we show the stabilizing effects of macroprudential policies in the case of economic or financial distress.

Although we performed our first simulations of the effects of macroprudential policies using the agent-based macro model in 2020, mostly the latest research results from the MACROPRU¹ project will be reported. The simulations were performed for the selected European Union countries using the European Central Bank's disaggregated data.

¹ This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101023445. The project MACROPRU: "Investigating new policies for financial stability that do not create inequality" has been carried out under the supervision of J. Doyne Farmer (Complexity Economics Group at the Institute for New Economic Thinking & Mathematical Institute) since October 2021.

Nearest-Neighbour Resampling for Off-Policy Policy Evaluation with Applications in Market Making

¹*Giegrich, M., ²Oomen, R., ¹Reisinger, C.

*lead presenter

¹Mathematical Institute Oxford University, United Kingdom

²Deutsche Bank, United Kingdom

In many real-world reinforcement learning problems, a lot of observational data is available while interaction with the original learning environment itself is costly (i.e. market making in limit order books (LOB)). Off-policy policy evaluation deals with the problem of estimating the value function of a target policy using data generated by a different (observed) policy. In this paper we propose a to-our-knowledge novel procedure for off-policy policy evaluation based on a nearest-neighbour resampling algorithm for deterministic policies in feedback form. Assuming continuity in the reward distribution and the state transition kernel, the procedure exploits that similar state/action pairs are associated with similar rewards and state transitions. This data-driven procedure has the advantage of being less wasteful with data compared to importance sampling if the observed policy has no or only a little mass on the deterministic target policy. Furthermore, the method does not explicitly assume a model such as in model-based off-policy policy evaluation. Understanding the algorithm as subsampling a regression estimator, we can analyse some of its properties and we conduct further numerical experiments.

We apply this algorithm to the problem of market making in pro-rata LOB markets providing a novel and data-driven way of generating synthetic LOB paths from a large scale data set of real observational LOB data. We investigate the quality of our simulations and evaluate the performance of certain trading policies. Since LOB data is often seen as non-Markovian we consider strategies to expand the state space with summary statistics and to avoid the curse of dimensionality in nearest-neighbour matching we investigate the usage of dimension reduction techniques and the choice of matching metric.

Statistical and agent-based modelling of the transmissibility of different SARS-CoV-2 variants in England and impact of different interventions

¹*Panovska-Griffiths, J., ²Swallow B., ¹Hinch R., ³Cohen J., ³Rosenfeld K,
⁴Stuart R.M., ¹Ferretti L., ¹Di Lauro F., ¹Wymant C., ³Izzo A., ⁵Waite W.,
⁶Viner R., ⁷Bonell C., ⁸The COVID-19 Genomics UK (COG-UK) Consortium,
¹Fraser C., ³Klein D., ³Kerr C.C.

*lead presenter

¹[https://www.cogconsortium.uk](#), University of Oxford, UK

²University of St. Andrews, UK

³Institute for Disease Modeling, Bill and Melinda Gates Foundation, USA

⁴University of Copenhagen, Denmark

⁵University of Strathclyde, UK

⁶University College London, UK

⁷London School of Hygiene and Tropical Medicine, UK

⁸<https://www.cogconsortium.uk>

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus causing COVID-19 spread in England throughout 2020 and 2021 facilitated by the emergence of new viral variants such as B.1.177, B.1.1.7 (Alpha) and B.1.617.2 (Delta), which dominated in late 2020 and early 2021. One of the models that offered ongoing and timely informed evidence on the effect of different strategies in the UK (throughout 2020) and England specifically (throughout 2021-2022) was Covasim (COVID-19 Agent-based Simulator). Covasim is an agent-based model developed by the Institute for Diseases Modelling at the Bill and Melinda Gates Foundation – initially formulated to explore the impact of contact tracing which is challenging to do properly with population-based SEIR models. Technical details of the model can be found in [1]. In early 2020, I had used the general Covasim framework equipped with UK and England specific data, to develop UK and England specific models to track these epidemics. This talk will illustrate applications of the modelling to give informed scientific advice at different policy junctions. We will focus on three case studies: a) schools reopening in September 2020 following the first epidemic wave (details in [2]), b) schools reopening at stage 1 of the Reopening Roadmap in March 2021 following the alpha wave (details in [3]) and c) reopening of society in June/July 2021 amidst the Delta wave at stage 4 of the Reopening Roadmap (details in [4]).

For the first case study we will show how school could safely reopen in September 2020 with effective Test-Trace-Isolate (TTI) strategy. For the second case study we will illustrate why schools had to be closed in January 2021 against the rising Alpha wave, but the conditions were right for them to reopen safely in March 2021. For the third study we will combine statistical models and Covasim to the progressive transmissibility of emerging variants and estimate B.1.177 to be 20% more transmissible than the wild type, Alpha to be 50-80% more transmissible than B.1.177 and Delta to be 65-90% more transmissible than Alpha. Using these estimates we will showcase that due to the high transmissibility of Delta, resurgence in infections driven by the Delta variant would not be prevented, but would be strongly reduced by delaying the relaxation of restrictions by one month and with continued vaccination.

The results presented here and published in [2-4], alongside those of other modelling groups, were used to scientifically advise the UK Government to reopen schools in September 2020 and March 2021, but to delay Step 4 of the Reopening Roadmap till July 19, 2021.

Title: Disentangling homophily, community structure and triadic closure in networks

Abstract:

One of the most typical properties of network data is the presence of homophily, i.e. the increased tendency of an edge to exist between two nodes if they share the same underlying characteristic, such as a social parameter, metabolic role, etc. More broadly, when the underlying similarity parameter is not specified a priori, the same homophily pattern is known as community structure. Another pervasive pattern encountered in the same kinds of networks is transitivity, i.e. the increased tendency of observing an edge between two nodes if they share a neighbor in common. Although these patterns are indicative of two distinct mechanisms of network formation, namely choice or constraint homophily and triadic closure, respectively, they are generically conflated in non-longitudinal data. This is because both processes can result in the same kinds of observation: 1. the preferred connection between nodes of the same kind can induce the presence of triangles involving similar nodes, and 2. the tendency of triangles to be formed can induce the formation of groups of nodes with a higher density of connections between them, when compared to the rest of the network. This conflation means we cannot reliably interpret the underlying mechanisms of network formation merely from the abundance of triangles or observed community structure in network data.

In this talk I present a solution to this problem, consisting in a principled method to disentangle homophily and community structure from triadic closure in network data. This is achieved by formulating a generative model that includes community structure in a first instance, and an iterated process of triadic closure in a second. Based on this model, we develop a Bayesian inference algorithm that is capable of identifying which edges are more likely to be due to community structure or triadic closure, in addition to the underlying community structure itself. As we show, this reconstruction yields a detailed interpretation of the underlying mechanisms of network formation, allowing us to identify macro-scale structures that emerge spontaneously from micro-scale higher-order interactions, and in this way we can separate them from inherently macro-scale structures. I show how the method can evade the detection of spurious communities caused solely by the formation of triangles in the network, and how it can improve the performance of link prediction when compared to the pure version of the model without triadic closure.

Title: Over-parametrization: Insights from solvable models
Lenka Zdeborova

Abstract:

This talk gives an overview of recent results in a line of theoretical work that started 3 decades ago in statistical physics. We will first discuss teacher-student setting of the generalized linear regression. We illustrate the presence of the interpolation peak for classification with ridge loss and its vanishing with regularization. We show that, in the spherical perceptron, the optimally regularized logistic regression approaches very closely the Bayes optimal accuracy. We contrast this with the non-convex case of phase retrieval where the canonical empirical risk minimization performs poorly compared to the Bayes-optimal error. We then move towards learning with hidden units and analyze double descent in learning with generic fixed features and any convex loss. The formulas we obtain are generic enough to describe the learning of the last layer of neural networks for realistic data and networks. Finally, for the phase retrieval, we are able to analyze gradient descent in the feature-learning regime of a two-layer neural network where we show that over-parametrization allows a considerable reduction of the sample complexity. Concretely, an over-parametrized neural network only needs twice the input dimension of samples, while non-overparametrized network needs constant times more, and kernel regression quadratically many samples in the input dimension.

Computing second-order points under equality constraints using Fletcher’s augmented Lagrangian

¹Florentin Goyens*, ²Armin Eftekhari, ³Nicolas Boumal

Oral presentation

* lead presenter

¹ University Paris-Dauphine, France

² University of Umeå, Sweden

³ EPFL, Switzerland.

Given a Euclidean space $\mathcal{E}$ with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|$, we consider a generic smooth nonconvex optimization problem with equality constraints

$$\min_x f(x) \text{ subject to } h(x) = 0, \quad (\text{P})$$

where $f: \mathcal{E} \rightarrow \mathbb{R}$ and $h: \mathcal{E} \rightarrow \mathbb{R}^m$ are smooth. The feasible set is denoted by $\mathcal{M} = \{x \in \mathcal{E} : h(x) = 0\}$ and is assumed to be a smooth manifold. One emerging motivation for Problem (P) is the minimization over Riemannian manifolds using infeasible methods, which is valuable when computing retractions is prohibitively expensive, as is the case with orthogonality constraints when the number of columns is large. There are indications that orthogonality is helpful to mitigate the effect of adversarial attacks against a neural network. Fletcher (1970) introduced the penalty function

$$g(x) := \mathcal{L}_\beta(x, \lambda(x)) = f(x) - \langle h(x), \lambda(x) \rangle + \frac{\beta}{2} \|h(x)\|^2$$

with a particular choice of multipliers $\lambda(x) \in \mathbb{R}^m$. We introduce a natural definition of approximate criticality for (P) at non-feasible points, which is based on the Riemannian gradient and Hessian of f on a manifold near the feasible set. We propose an algorithm based on the minimization of Fletcher’s augmented Lagrangian which produces an $(\varepsilon_1, 2\varepsilon_1)$ –first-order critical of (P) in at most $\mathcal{O}(\varepsilon_1^{-2})$ iterations and an $(\varepsilon_1, 2\varepsilon_1, \varepsilon_2 + C\varepsilon_1)$ –second-order critical point of (P) in at most $\mathcal{O}(\max\{\varepsilon_1^{-2}, \varepsilon_2^{-3}\})$ iterations, where $C \geq 0$ is a constant depending on the function h . This matches the optimal worst-case complexity with respect to ε , and is the first infeasible algorithm to achieve such rate in this setting for a similar notion of approximate optimality.

References

Fletcher, R. (1970). A class of methods for nonlinear programming with termination and convergence properties. Integer and nonlinear programming, pages 157–173.

Higher-order signal processing with the Dirac operator

^{1*}Lucille Calmon, ²Michael Schaub, ^{1,3}Ginestra Bianconi

*lead presenter: oral presentation in person,

¹Queen Mary University of London, United Kingdom

²RWTH Aachen University, Germany

³Turing Institute, United Kingdom

Graph signal processing has proven essential to the filtering and analysis of real data supported on networks. However, many real systems contain higher-order interactions which cannot be simply represented by the pairwise paradigm of graphs. Consequently, there has been increasing interest to investigate frameworks that encode higher-order interactions and dynamics, such as simplicial complexes.

Topological signals are variables not only supported on the nodes, but also links and higher order simplices in a simplicial complexes. They can for example encode fluxes and edge flows in dynamical systems.

While the Hodge-Laplacian of simplicial complexes has provided an appropriate shift operator to extend the formalism of signal processing to topological signals, this shift operator treats all topological signals of different orders independently. There remains a strong need to develop a framework to treat topological signals of different orders together, as they may arise in data.

Here, we propose a novel shift operator based on the recently proposed Dirac operator of simplicial complexes. Exploiting the spectral properties of the Dirac operator and its relation to the Hodge-Laplacian, we show how it provides an elegant easily generalisable framework to conduct a coupled processing of signals defined on nodes, links, triangles etc. The Dirac operator topologically couples the different signals together, which in turn allows conjunct processing at different orders. We demonstrate how this enables us, e.g., to leverage node signals for the processing of edge flows.

In this talk we study the convergence properties (both to a first order and second order critical point) of a subspace cubically regularised method for optimisation of large-scale non-convex smooth objectives. Cubical regularisation-based method for non-convex smooth objectives has been shown to be convergent to a first order and second order critical point, whose function/gradient evaluation complexity (for first order critical point) is an order of magnitude smaller than the traditional trust-region or line search/gradient-descent methods. We show theoretically that replacing the full variable space with a subspace generated by structured random matrices gives a method that retains the convergent properties and the fast convergent rate in terms of the iteration complexity compared to the full-space version, when the Hessian of the objective has a low-rank. The reduced dimension demands less memory requirement and computational resources when the objective has a large variable space, e.g., in the training of neural networks

Spectral Tools for training and analysing Neural Networks

^{1,2}Lorenzo Giambagli et. al

¹ [_____](#), CSDC, Department of Physics and Astronomy, University of Florence, Sesto Fiorentino, Italy

²naXys - Namur Centre for Complex Systems, University of Namur, Namur, Belgium

Deep Feedforward Neural Networks (FFNNs) play a central role in the Machine Learning field. They are usually trained in the space of nodes, by adjusting the weights of existing links via suitable optimization protocols. Recently a radically new approach has been proposed (1). By anchoring the learning process to reciprocal space, the new targets of the optimization process are eigenvectors and eigenvalues of the transfer operators between layers.

Shifting the focus on such fundamental mathematical structures we have been able to understand their pivotal role in training and analyzing NNs. Indeed, while seeking for a small subset of trainable parameters capable of carrying out the training procedure, eigenvalues are what to look for (2). Choosing them as trainable parameters allows the optimizer to exploit the parallel adjustment of several weights, the ones underlined by the corresponding eigenvector, and therefore made their after-training interpretation possible.

Firstly, eigenvalues magnitude after the training procedure has occurred has been empirically and heuristically proven being a proxy their relevance in the optimization process. Indeed, a precise correspondence between nodes and eigenvalues can be established, leading to a novel pruning procedure. The nodes related with low magnitude eigenvalues can be removed leading to a fast and easy implemented network compression algorithm (3), Fig.1.

Secondly, accounting for eigenvalues in the optimization process, it is possible to dynamically train sparse network (2). Sparsity constrains in the direct space implies that certain weights got filtered under a mask, leading to a gradient equal to zero during the training procedure. Working in the reciprocal space, however, allows masked weights to still be modified, due to the non-local effect of the eigenvalues. Such approach leads to sparse networks whose topology is not fixed to the starting one, resulting in a much more efficient training.

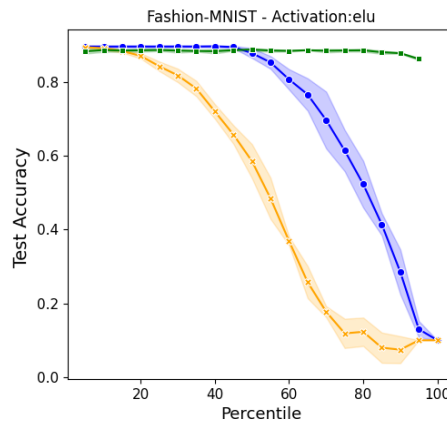


Figure 1: Accuracy drop with respect to percentile of nodes removed while eigenvalues are employed as a pre- (green) post- (blue) training pruning. In yellow the curve for the conventional case where connectivity is used instead as a proxy for relevance.

1. *Machine learning in spectral domain*. **Lorenzo Giambagli, Lorenzo Buffoni, Timoteo Carletti, Walter Nocentini & Duccio Fanelli**. 2021, Nature Communications.
2. *On the training of sparse and dense deep neural networks: less parameters, same performance*. **Lorenzo Giambagli, Lorenzo Chicchi, Lorenzo Buffoni, Timoteo Carletti, Marco Ciavarella, Duccio Fanelli**. 2021, PRE.
3. *Spectral Pruning of Fully Connected Layers*. **Lorenzo Giambagli, Lorenzo Buffoni, Enrico Civitelli, Lorenzo Chicchi, and Duccio Fanelli**. 2021, (under peer review of Scientific Report).

The success of modern deep learning methods reveals unarguably that standard supervised learning tasks, such as image classification, are very special points in the space of all possible objective functions. A characterisation of such points is then instrumental to the theoretical understanding of deep learning methods and their generalisation capabilities. In other words, what kind of target functions should the theory of deep learning be concerned with? We have recently proposed an idealised learning scenario where to answer this question. This scenario is based on generating objective functions with a controlled structure: for instance, one can think of large-dimensional functions which are linear combination of constituent functions of smaller dimensionality, not unlike images which are made by gluing parts which display different features. The objective is then learnt by very wide neural networks operating in the lazy regime, where the equivalence with kernel methods allows for a precise characterisation of the generalisation capabilities. As a practical example, one can use a scenario of this kind prove the benefits of convolutional neural networks over fully-connected ones in learning local functions. In this talk I will illustrate our efforts in generalising this idea to objective functions of increasing complexity, inspired by the structure of natural images. In fact, natural images are believed to possess a certain compositional structure, whereby features at any given scale are made of sub-features of smaller scale. Therefore, the set of relevant objective functions must include linear combinations of constituent functions of smaller dimensionality as well as iterative compositions of these constituents. Do deep convolutional networks have an edge over other architectures on such functions?

Accelerating Deep Unrolling Networks via Dimensionality Reduction

Junqi Tang, Subhadip Mukherjee, Carola-Bibiane Schönlieb

JT

University of Cambridge, UK

In this work we propose a new paradigm for designing efficient deep unrolling networks using stochastic approximation and operator sketching for dimensionality reduction. The deep unrolling networks are currently the state-of-the-art solutions for imaging inverse problems. However, for high-dimensional imaging tasks, especially the 3D cone-beam X-ray CT and 4D MRI imaging, the deep unrolling schemes typically become inefficient both in terms of memory and computation, due to the need of computing multiple times the high-dimensional forward and adjoint operators.

We first found that such limitations can be partially addressed by stochastic unrolling with subsets of operators, inspired by the success of stochastic first-order optimization. We develop a stochastic (ordered-subsets) variant of the classical learned primal-dual (LPD) (Adler and Öktem, 2018), which is a state-of-the-art unrolling network for tomographic image reconstruction. The proposed learned stochastic primal-dual (LSPD) network (Tang et al., 2021) only uses subsets of the forward and adjoint operators and offers considerable computational efficiency. We provide theoretical analysis of a special case of our LSPD framework, suggesting that it has the potential to achieve image reconstruction quality competitive with the full-batch LPD while requiring only a fraction of the computation. The numerical results for two different X-ray computed tomography (CT) imaging tasks (namely, low-dose and sparse-view CT) corroborate this theoretical finding, demonstrating the promise of LSPD networks for large-scale imaging problems.

We then propose a further acceleration upon stochastic primal-dual unrolling, using sketching techniques to approximate products in the high-dimensional image space (Tang et al., 2022). The operator sketching can be jointly applied with stochastic unrolling for the best acceleration and compression performance. Our numerical experiments on X-ray CT image reconstruction demonstrate the remarkable effectiveness of our sketched unrolling schemes.

References

- Jonas Adler and Ozan Öktem. Learned primal-dual reconstruction. *IEEE transactions on medical imaging*, 37(6):1322–1332, 2018.
- Junqi Tang, Subhadip Mukherjee, and Carola-Bibiane Schönlieb. Stochastic primal-dual deep unrolling. *arXiv preprint arXiv:2110.10093*, 2021.
- Junqi Tang, Subhadip Mukherjee, and Carola-Bibiane Schönlieb. Operator sketching for deep unrolling networks. *arXiv preprint arXiv:2203.11156*, 2022.

Title: Approximation properties of two-layer neural networks with values in a Banach space

Abstract: Approximation properties of infinitely wide neural networks have been studied by several authors in the last few years. New function spaces have been introduced that consist of functions that can be efficiently (i.e., with dimension-independent rates) approximated by neural networks of finite width. Typically, these functions are supposed to act between Euclidean spaces, typically with a high-dimensional input space and a lower-dimensional output space. As neural networks gain popularity in inherently infinite-dimensional settings such as inverse problems and imaging, it becomes necessary to analyse the properties of neural networks as nonlinear operators acting between infinite-dimensional spaces. In this talk, I will present dimension-independent Monte-Carlo rates for neural networks acting between Banach spaces with a partial order (vector lattices), where the ReLU non-linearity will be interpreted as the lattice operation of taking the positive part.

Identifiability in Inverse Reinforcement Learning

¹Cao, H., ^{*2}Cohen, S.N., & ³Szpruch, L.

¹ _____, Ecole Polytechnique, France

² _____, Mathematical Institute, Oxford and Alan Turing Institute

³ _____, School of mathematics, University of Edinburgh and Alan Turing Institute

Inverse reinforcement learning attempts to reconstruct the reward function in a Markov decision problem, using observations of agent actions. This problem is ill-posed, and the reward function is not identifiable, even under the presence of perfect information about optimal behavior. We provide a resolution to this non-identifiability for problems with entropy regularization, and with some uncertainty averse agents.

For a given discrete-time finite-state MDP, we fully characterize the reward functions leading to a given policy and demonstrate that, given demonstrations of actions for the same reward under two distinct discount factors, or under sufficiently different environments, the unobserved reward can be recovered up to a constant. We also give general necessary and sufficient conditions for reconstruction of time-homogeneous rewards on finite horizons, and for action-independent rewards, generalizing recent results of Kim et al. [2021] and Fu et al. [2018]

An unsupervised manifold learning approach to untangle big data of collective human behaviors

¹*Fanqi Zeng, ¹Nikolai Bode, ¹Martin Homer and ²Thilo Gross

*lead presenter

¹Department of Engineering Mathematics, University of Bristol, Bristol, UK

²Helmholtz Institute for Functional Marine Biodiversity, University of Oldenburg, Oldenburg, Germany

Collective human behaviors, such as crowd stampedes in confined spaces, are complicated to analyze because of many nonlinear dynamics and variations. In this work, we propose using the diffusion map, an unsupervised manifold learning method, to identify the hidden patterns in collective pedestrian behaviors. We build a feature space by defining a set of measurements to extract observables from pedestrians' trajectory recordings. By applying the diffusion map on the feature spaces of one experimental data set and three simulated data sets of crowd evacuations, we then obtain the leading eigenvectors containing enriched data information. We recover the speed-density relationship from the data and distinguish different data sources based on the leading eigenvectors. We also identify the abnormal individuals without checking the original recordings. Our work contributes a convenient approach to analyzing collective human behavior without needing much prior knowledge of the data .

Phase Collapse in Neural Networks

^{1,2}*Guth, F., ²Zarka, J. & ^{3,4}Mallat, S.

*lead presenter

¹

²DI, ENS, CNRS, PSL University, Paris, France

³Collège de France, Paris, France

⁴Flatiron Institute, New York, USA

Deep convolutional classifiers linearly separate image classes and improve accuracy as depth increases. They progressively reduce the spatial dimension whereas the number of channels grows with depth. Spatial variability is therefore transformed into variability along channels. A fundamental challenge is to understand the role of non-linearities together with convolutional filters in this transformation.

ReLUs with biases are often interpreted as thresholding operators that improve discrimination through sparsity. We demonstrate that it is a different mechanism called phase collapse which eliminates spatial variability while linearly separating classes. We show that collapsing the phases of complex wavelet coefficients is sufficient to reach the classification accuracy of ResNets of similar depths.

However, replacing the phase collapses with thresholding operators that enforce sparsity considerably degrades the performance. We explain these numerical results by showing that the iteration of phase collapses progressively improves separation of classes, as opposed to thresholding non-linearities.

I (F. Guth) intend to present this work in person. It is based on the following paper: F. Guth, J. Zarka, and S. Mallat. Phase Collapse in Neural Networks. In *International Conference on Learning Representations, ICLR*, 2022.

Sequential data assimilation for uncovering an influence network from Hawkes processes driven by count data

¹*Santitissadeekorn, Naratip, ¹Lloyd, David., ²Short, Martin.

*lead presenter

1 Department of Mathematics, University of Surrey, UK

2 Department of Mathematics, Georgia Institute of Technology, USA

This talk will describe a novel method for constructing an influence network from a time series of spatiotemporal count data, which encompasses an arbitrary number of events in a time interval, both as a batch and real-time data stream. The intensity of the count data time series will be modelled by the (parametric) multivariate Hawkes process.

A new efficient filtering algorithm estimates and tracks parameters of the Hawkes process model. The method uses information from sequences of count data observed in a fixed time interval for different stochastic processes to discover a directed network whose links represent the influence among the nodes. The challenge to achieve the feasibility of the method for big data requires some simplification and assumption of the algorithm to enable parallelisation. This work will also expand the capability of the existing cutting-edge inference methods for influence networks with underlying Hawkes processes, which are mostly limited only to the timestamp data but not the times-series of count data. A potential breakthrough of this work will remove this limitation and expand the utility of the stochastic point-process framework that can assimilate a sequence of count data for an online inference and uncertainty quantification of influence networks. From a practical standpoint, the research work will also provide an uncertainty analysis of the inferred network structure. The outcome of this work will also facilitate downstream uncertainty analysis for network applications such as node classifications, link prediction and rare-event detection.

GNNs for Node Clustering in Signed and Directed Networks

^{1*}Yixuan He ^{1,2}Gesine Reinert ^{1,2}Mihai Cucuringu

*lead presenter

¹ [_____](#), University of Oxford, United Kingdom

²The Alan Turing Institute, United Kingdom

With an increasing number of applications where data can be represented as graphs, graph neural networks are a useful tool to apply deep learning to graph data. In particular, node clustering is an important problem in network analysis. Signed networks are ubiquitous in many real-world applications (e.g., social networks encoding trust/distrust relationships, correlation networks arising from time series data), among which many networks are themselves directed. While most existing methods that could be applied to directed clustering use local edge densities as main signal and directionality as additional signal, we argue that even in the absence of any edge density differences, directionality can play a vital role in directed clustering as it can reveal latent properties of network flows. Signed and directed networks are important types of networks that are linked to many real-world problems; their asymmetry provides a challenge for many clustering methods.

We propose two graph neural network models for node clustering in signed networks and directed networks, respectively. The methods are end-to-end in combining embedding generation and clustering without an intermediate step. Experimental results on a synthetic signed stochastic block model, a polarized version of it, and real-world data at different scales, demonstrate that our proposed methods can achieve comparable or better results than state-of-the-art node clustering methods, for a wide range of noise and sparsity levels. The introduced models complement existing well-performing methods through the possibility of including exogenous information, in the form of node-level features or labels.

Motion correction for cardiac MRI images using generative neural networks with disentangled latent space variables

^{1,2} Abdullah, ² Martin Holler, ² Karl Kunisch, ^{3*} Malena Sabaté Landman

¹ Institute of Mathematics and Scientific Computing, University of Graz, Austria

² Department of Mathematics, the Chinese University of Hong Kong, Hong Kong

³ [_____](#), Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK

Patient motion poses a great challenge in magnetic resonance imaging (MRI) due to its effects on the image reconstruction as well as for its subsequent image interpretation. Although dynamic MRI aims to image particular physiological dynamics, for example heart motion, dynamic images might also be distorted by additional motion artifacts, such as those arising from breathing motion. The focus of this work is to remove such distortions while maintaining the desired dynamics. To this aim, a method based on generative neural networks is proposed, that represents an image sequence as the forward mapping via a generative neural network of latent space variables such that different sources of motion are characterized independently via latent space disentanglement using one dimensional prior information. This learned representation allows to freeze one (unwanted) type of motion while maintaining the relevant dynamics. The proposed learning-free method is fully unsupervised as all the network parameters are learned directly from the given data. Subsequently, this method is not susceptible to bias coming from a training dataset because it does not require any prior training. The performance of the algorithm is shown both on synthetic and real dynamic MRI data that display two physically different motion types, breathing and heart-beat motion, where we can successfully disentangle the different motions in the latent space and use that to freeze the breathing motion while maintaining an accurate representation of the heart-beat motion.

Mirror Q-learning can exploit the structure of the environment

¹*Alfano, C., ¹Rebeschini, P.

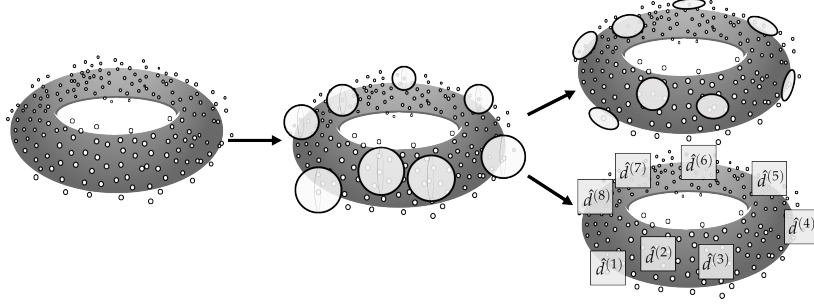
*lead presenter

¹[University of Oxford](#), United Kingdom

In this work, we apply the mirror descent algorithm to Q-learning and study how to take advantage of its desirable properties in the Reinforcement Learning (RL) framework. Q-learning is one of the main paradigms of RL and consists in learning the optimal Q-function of a Markov decision process (MDP) in a model-free fashion. In recent works, the structure of the MDP and of the Q-function has received a lot of attention, as researchers have exploited certain properties, for instance low-rankedness or linearity, to achieve better sample complexity or to learn an efficient representation of the environment by designing tailored algorithms. However, these algorithms often require the environment properties to be explicit, which is not always possible. As mirror descent has already been shown to be able to implicitly take advantage of data structures, our objective is to show that it can be applied to Q-learning in order to formulate a general and fundamental algorithm that can still recover the results of previous works in an implicit fashion. We prove that, for a wide family of mirror maps, our version of mirror Q-learning achieves the same sample complexity as classical Q-learning. Additionally, we demonstrate that a particular choice of a mirror map can take advantage of the structure of the Q-function. We make a case study out of the low-rank Q-function setting, where we show that choosing the hyperentropy mirror map results in the algorithm converging to a low-rank solution and in sample complexity savings. Finally, we corroborate our results by showing the performance of our algorithm on several stochastic control tasks.

TANGENT SPACE AND DIMENSION ESTIMATION WITH THE WASSERSTEIN DISTANCE

UZU LIM, HARALD OBERHAUSER, VIDIT NANDA



Preprint - <https://arxiv.org/abs/2110.06357>

Abstract. The *manifold hypothesis* is a conventional assumption in data analysis: a given dataset is assumed to be sampled from a d -dimensional data manifold $M \subseteq \mathbb{R}^D$. While heuristic analyses based on the manifold hypothesis are plentiful, precise mathematical theorems are scarce.

We provide mathematically rigorous theorems that guarantee correct inference of two important properties of the data manifold - its tangent spaces and its intrinsic dimension d . In data science terms, a tangent space gives the optimal local linear regression, and the intrinsic dimension is the degree of freedom of data. The algorithm for this inference is a local version of the principal components analysis (PCA). To our knowledge, our results are the first ever to enjoy the following properties - (1) accommodating non-uniformity of the data distribution, (2) accommodating non-uniform noise that may vary across the manifold, and (3) explicit computation of all of the constants appearing in the theorem.

Our results translate an intuitive reasoning to precise mathematics involving random matrix theory and optimal transport theory. The key intuition is that every manifold is approximately linear when examined locally, and that PCA provides optimal linear approximations. Our argument proceeds as follows. We first prove a local version of the matrix Hoeffding's inequality to understand the convergence rate of local empirical covariance matrices. Then we use the Wasserstein distance to quantify how non-linear a manifold gets when examined locally. Finally, simple applications of the Davis-Kahan theorem and the Hoffman-Wielandt theorem separately yield our 2 main theorems.

UZU LIM, HARALD OBERHAUSER, VIDIT NANDA

This presentation is intended to happen in person.

Authors' address: Mathematical Institute, University of Oxford, Radcliffe Observatory, Andrew Wiles Building, Woodstock Rd, Oxford OX2 6GG

Unifying information propagation models and their influence maximisation

^{1*}Y. Tian, ²R. Lambiotte.

¹Mathematical Institute, University of Oxford, Oxford, UK, [_____](#).

²Mathematical Institute, University of Oxford, Oxford, UK

Information propagation is a central theme in social, behavioural, and economic sciences, with important theoretical and practical implications, such as the influence maximisation problem for viral marketing. There are two widely adopted models in this context: the independent cascade model where nodes adopt their behaviour from each neighbour independently, and the linear threshold model where collective effort from the whole neighbourhood is needed to influence a node. However, both models suffer from certain drawbacks, including a binary state space, where nodes are either active or not, and the absence of feedback, as nodes can not be influenced after having been activated previously. To address these issues, we consider a model with continuous variables that has the additional advantage of unifying the mechanisms underlying these two classic models. For the associated influence maximisation problem, the objective function is no longer submodular, a feature that most approximation algorithms are based on but is arguably strict in practice. Hence, we develop a framework, where we formulate the influence maximisation as a mixed integer nonlinear programming and adopt derivative-free methods as general solutions. Furthermore, we show that the problem can be exactly solved in the special case of linear dynamics, and propose a customised direct search method accordingly, with local convergence. We demonstrate the rich behaviour of the newly proposed information propagation model and the close-to-optimal performance of the customised direct search method numerically on both synthetic and real networks.

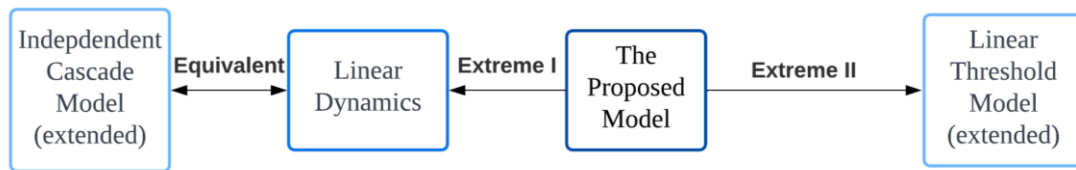


Figure 1: An illustration of the proposed general class of information propagation model, where the classic independent cascade model and linear threshold model after extended for continuous variables and deterministic dynamics can be recovered by setting appropriate model parameters.

References

[1] Y. Tian, R. Lambiotte, *arXiv:2112.01465* (2021).

Using Web Data and Network Science to Detect Spatial Relationships

^{1*}Dilworth, ²E., Tranos, E. & ¹Lawson, D.

¹ , [School of Mathematics, University of Bristol, Bristol, UK.](#)

²School of Geographical Sciences, University of Bristol, Bristol, UK; and The Alan Turing Institute, London, UK.

Spatial relationships are at the core of spatial sciences, however geographers often lack means of observing spatial relationships. Understanding how two places are connected is useful, but can only be inferred and not observed. Being able to capture trade and connections between regions, and how these change with time, are spatial relationships that currently there is a limited understanding of.

Using data from the Internet Archive's web collection related to the UK, which has been collected over the period 1996-2013, hyperlinks between websites over time and the geolocations of the websites are recorded, providing data with both spatial and temporal qualities.

By aggregating the data to a level that avoids too high sparsity, methods for analysing embedded dynamic networks can be applied. The data can be put into undirected and directed graphs, to develop an understanding of the spatial locations and volumes of which there are website connections in the UK.

By viewing the data as a sequence of networks we can "embed" it into a lower dimensional space. Doing this allows new patterns to be observed and allows spatio-temporal clusters of regions with similar connection profiles to be formed.

A joint spectral embedding of multiple graphs technique called *unfolded adjacency spectral embedding (UASE)* is a process that can be applied to the data described, to see how the graphs evolve with time. It also helps us to identify patterns in spatial relationships. The data being used in this analysis throws up problems due to degree heterogeneity, which can be addressed with degree correction methods.

The findings can be interpreted to understand the changes in behaviour of websites, communities, or the entire graph, that happen with time. We can visualise where clusters of communities are linked with one another are located spatially, and how these connections evolve with time.

Sequential Motifs in Observed Walks Through Networks

^{1*} LaRock, T., ^{2,3} Scholtes, I, ^{4,5} Eliassi-Rad, T.

¹ [_____](#), Mathematical Institute, University of Oxford, UK

² Data Analytics Group, Department of Informatics, University of Zürich, Switzerland

³ Chair of Machine Learning for Complex Networks, Center for Artificial Intelligence and Data Science, Julius-Maximilians-Universität, Germany

⁴ Network Science Institute, Northeastern University, USA

⁵ Khoury College of Computer Sciences, Northeastern University, USA

The structure of complex networks can be characterized by counting and analyzing network motifs. Motifs are small graph structures that occur repeatedly in a network, such as triangles or chains. Recent work has generalized motifs to temporal and dynamic network data. However, existing techniques do not generalize to sequential or trajectory data, which represents entities moving through the nodes of a network, such as passengers moving through transportation networks. The unit of observation in these data is fundamentally different, since we analyze observations of trajectories (e.g., a trip from airport A to airport C through airport B), rather than independent observations of edges or snapshots of graphs over time. In our recent work,[†] we define sequential motifs in trajectory data, which are small, directed, and sequenced-ordered graphs corresponding to patterns in observed sequences. We draw a connection between counting and analysis of sequential motifs and Higher-Order Network (HON) models. We show that by mapping edges of a HON, specifically a k th-order DeBruijn graph, to sequential motifs, we can count and evaluate their importance in observed data. We test our methodology with two datasets: (1) passengers navigating an airport network and (2) people navigating the Wikipedia article network. We find that the most prevalent and important sequential motifs correspond to intuitive patterns of traversal in the real systems and show empirically that the heterogeneity of edge weights in an observed higher-order DeBruijn graph has implications for the distributions of sequential motifs we expect to see across our null models.

References

[†] LaRock, T., Scholtes, I, Eliassi-Rad, T. Sequential Motifs in Observed Walks. *arXiv preprint arXiv:2112.05642* (2021).

Robust Dynamic Network Embedding

Ed Davis

[University of Bristol](#), United Kingdom

Here, I visit the problem of dynamic network embedding, that is, representing the nodes of a dynamic network as evolving vectors within a low-dimensional space. Dynamic networks encode both spatial and temporal relations between nodes, relations which are desirable for a large range of applications from cyber security to neuroscience. Representing the network in such a space allows us to detect spatial and temporal patterns through clustering. In a recent paper, Gallagher et al. [1] showed that unfolded adjacency spectral embedding (UASE) provides a theoretically grounded procedure for obtaining a dynamic network embedding which is stable in both time and space, allowing dynamic community structure to be recovered consistently using spatiotemporal clustering. However, the authors leave the question open of how to achieve such guarantees with more robust procedures, for example, analogous to normalized or regularised Laplacian spectral embedding in the simple graph case. In this talk, I will present research on this question, providing a practical procedure which is scalable and robust to common real-world network properties such as sparsity, degree heterogeneity, and directedness of graphs. Amongst other practical applications, we find that our approach correctly identifies school children in the same class and year in a temporal contact graph. Importantly the procedure remains correct in the context of degree correction and scales to very large datasets.

[1] I. Gallagher, A. Jones, and P. Rubin-Delanchy, “Spectral embedding for dynamic networks with stability guarantees,” arXiv preprint arXiv:2106.01282, 2021.

Title: The transformer leap

Nandos de Freitas

Abstract: In the last five years we witnessed an extraordinary leap in our understanding of intelligence and the capabilities of AI systems. Large scale training of transformer neural networks was the driving force, but it soon became apparent that very simple supervisory signals could enable us to train foundations models, which can be easily adapted to solve a myriad of problems. Transformers have taken over, solving protein folding, generating language, generating images and videos to illustrate a sentence, mapping natural language descriptions to code, and ultimately paving the way for the design of intelligent generalist agents. In this talk, I will share a personal view of this transformer leap.