The Abdus Salam International Centre for Theoretical Physics

United Nations Educational, Scientific and Cultural Organization

IAEA International Atomic Energy Agency

Youth in High-Dimensions: Recent Progress in Machine Learning, High-Dimensional Statistics and Inference | (SMR 3841)

29 May 2023 - 02 Jun 2023
ICTP, Trieste, Italy

P01 - ADEBAYO Segun

Application of suitable statistical tool for effective inference in a multimodal dataset scenario

P02 - ADENOMON Monday Osagie

Predicting Wheaton Precious Metal Stock Prices: Machine Learning Techniques versus Time Series Approach

P03 - ADOMAITYTE Urte

Classification via empirical risk minimisation of power-law distributed clouds

P04 - AIUDI Riccardo

Predicting the effect of training on the internal representations of finite-width Bayesian one-hidden layer networks

P05 - ANNESI Brandon Livio

A sunburst of solutions in the continuous negative perceptron

P06 - ARIOSTO Sebastiano

Statistical mechanics of deep learning beyond the infinite-width limit.

P07 - ARPINO Gabriel

Statistical-Computational Tradeoffs in Mixed Sparse Linear Regression

P08 - AWORINDE Oluwatobi Halleluyah

Employing Deep Learning Model for Malar Region-Based Groupwise Age Ranking

P09 - BALLARIN Emanuele

CARSO: CounterAdversarial Recall of Synthetic Observations

P10 - BARDONE Lorenzo

Beyond Gaussian models: a mathematical framework for the analysis of real-world data structures

P11 - BASILE Lorenzo

Relating Implicit Fourier Bias and Adversarial Attacks through Intrinsic Dimension

P12 - BOFFI Mathew Nicholas

Probability flow solution of the Fokker-Planck equation

P13 - BOMBARI Simone

Beyond the Universal Law of Robustness: Sharper Laws for Random Features and Neural Tangent Kernels

P14 - CATANIA Giovanni

Thermodynamics of attractor neural networks: from bi-directional to models with coupled-replicas

P15 - CROTTI Stefano

Large deviations in stochastic dynamics over graphs through Matrix Product Belief Propagation

P16 - CUI Chao Hugo

Learning (with) Deep Random Networks of Extensive Width

P17 - DEL TATTO Vittorio

Robust inference of causal links in high-dimensional dynamical processes from the information imbalance of rank statistics

P18 - DMITRIEV Daniil

Deterministic equivalent and error universality of deep random features learning

P19 - DURANTHON Odilon

Neural-prior stochastic block model

P20 - EPPING Bastian

Unified field theoretical approach to deep and recurrent neural networks

P21 - ERBA Vittorio

Statistical mechanics of the maximum-average submatrix problem

P22 - EVEN Mathieu

(S)GD over Diagonal Linear Networks: Implicit Regularisation, Large Stepsizes and Edge of Stability

P23 - FOLCHINI Sara

DFA for Continual Learning

P24 - FRANCAZI Emanuele

Understanding the effects of class imbalance on the dynamics of supervised learning

P25 - GIAMBAGLI Lorenzo

Spectral L2 Regularization of Feedforward Neural Networks

P26 - GRENIOUX Louis

Assisted Learning of Energy-Based Models with Normalizing Flows

P27 - GUPTA Ankita

Interplay of two finite reservoirs in a bidirectional transport system

P28 - HUANG Brice

Algorithmic Threshold for Multi-Species Spherical Spin Glasses

P29 - KENT-DOBIAS Patrick Jaron

How to count in hierarchical landscapes: quenched complexity for the spherical models

P30 - KÖGLER Kevin

Fundamental Limits of Two-layer Autoencoders, and Achieving Them with Gradient Methods

P31 - LIAN Peng

Impacts of Central-Pacific El Niño and physical drivers on Eastern Pacific tuna using Explainable Artificial Intelligence

P32 - LMAKRI Aziz

Statistical Inference for Multivariate Bilinear Panel Data

P33 - LOULA Joao

Large population models: scalable generative modeling from multiple non-representative databases

P34 - LUCIBELLO Carlo

The exponential capacity of modern associative memories

P35 - MACHADO PEREZ David

Evolution in time of the dynamical trajectories of discrete spin systems

P36 - MAMADOU Ndiaye

Nonparametric prediction and supervised classification for spatial dependent functional data under fixed sampling design.

P37 - MARGIOTTA Riccardo Giuseppe

Attacks on Online Learners: a Teacher-Student Analysis

P38 - MARIANI Matteo

Inference in conditioned dynamics through causality restoration

P39 - MELE Margherita

Exploring the weight space of a perceptron via enhanced sampling techniques

P40 - MESTRY Vasudev Dipali

Evaluating the risk of extinction of fish populations in natural habitat using Reversible-Jump Markov chain Monte Carlo method: A case study on the decline of Notopterus Chitala in India

P41 - MONTASSER Omar

Adversarially Robust Learning: A Generic Minimax Optimal Learner and Characterization

P42 - NEGRI Matteo

The Hidden-Manifold Hopfield Model and a learning phase transition

P43 - NYAKUNDI Nyatuga Gideon

Class-Imbalanced Classifiers: An Application in High Dimensional Datasets in Oncology

P44 - OKAJIMA Koki

Average case analysis of Lasso regression under ultra-sparse conditions

P45 - OLSEN Valdemar Lee Kargaard

Evaluating the quality of pairwise maximum entropy models in large neural datasets

P46 - PACELLI Rosalba

Statistical mechanics of deep learning beyond the infinite-width limit

P47 - PICCIOLI Giovanni

New tools for sampling the posterior of neural networks

P48 - REMIZOVA Anastasia

Analysis of test error in the asymptotic polynomial regime for two-layer neural networks with respect to activation function

P49 - RENDE Riccardo

Optimal inference of a generalised Potts model by a single-layer transformer with factored attention

P50 - ROBNIK Jakob

Microcanonical Hamiltonian Monte Carlo

P51 - SAWAYA Kazuma

A Unifying Approach for Statistical Inference in High-Dimensional Generalized Linear Models

P52 - SCAGLIOTTI Alessandro

AutoencODEs: a control theoretic framework for modeling Autoencoders

P53 - STEPHAN Théo Ludovic

From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks

P54 - SZEKELY Eszter

Data-driven separation between feature and lazy learners for higher-order statistics

P55 - THÉRIAULT Robin

The Amount of Data Needed to Train Dense Hopfield Networks

P56 - TIEPLOVA Daria

On the largest canonical correlation coefficients between the past and the future of a high-dimensional time series

P57 - TOMASINI Maria Umberto

How deep convolutional neural networks lose spatial information with training

P58 - TOSATO Niccolo

Emergent representations in networks trained with the Forward Forward algorithm

P59 - UDOMBOSO Godwin Christopher

On the Leveraging of MLab Facilities to Mine the Nigerian Internet Traffic Measurement

P60 - VALERIANI Lucrezia

The geometry of hidden representations of large transformer models

P61 - VELASQUEZ VARELA Diego

An Application of Correlation-Based Clustering for data imputation

P62 - VENTURA Enrico

Training neural networks with structured noise improves classification and generalization

P63 - VITERITTI Luciano Loris

Transformer variational wave functions for frustrated quantum spin systems

P64 - WANG Jie Yu

Study on thermodynamics and kinetics of nucleic acid base pairs by statistical physics

P65 - WANG Lele

Attributed Graph Alignment: Fundamental Limits and Efficient Algorithms

P01

Application of suitable statistical tool for effective inference in a multimodal dataset scenario

Abstract

Raw data in general is unlikely to reveal interesting patterns because human perceptual abilities, although powerful, only deal with clear patterns of surface contrasts, such as perceptible changes in contrast or rhythm. Moreover, with the era of big data, obtaining meaningful inference from dataset without the application of basic statistical techniques is becoming seemingly impossible. Improving awareness of the use of statistical tool is particularly important at this time because dataset now faces the challenge of moving on from a period of broad theory construction to studies that build on those foundations. Reliably locating significant patterns in data is a crucial precondition for such work and statistical methods provide established and well-understood techniques for just this task. However, most classical statistical methods are designed for use on low-dimensional data. These are data whose number of observations n is much larger than the number of features p. In the exploration of low-dimensional datasets, it is possible to plot the response variable against each of the limited number of explanatory variables to get an idea which of these are important predictors of the response. However, with high-dimensional data such as multimodal, the large number of explanatory variables makes this process difficult. In some high-dimensional datasets it can also be difficult to identify a single response variable, making standard data exploration and analysis techniques less useful. The study explores four types of statistical techniques capable of analyzing high dimensional dataset: regression with numerous outcome variables, regularized regression, dimensionality reduction, and clustering. The strength and challenges of each of these techniques were considered.

P02

Predicting Wheaton Precious Metal Stock Prices: Machine Learning Techniques versus Time Series Approach

[*1,2,3]Adenomon, M. O

1. Department of Statistics, Nasarawa State University, Keffi, Nigeria & NSUK-LISA Stat Lab, Nasarawa State University, Keffi, Nigeria
2. Chair, International Association of Statistical Computing (IASC) African Members Group
3. Foundation of Laboratory for Econometrics and Applied Statistics of Nigeria (FOUND-LEAS-IN-NIGERIA)

adenomonmo@nsuk.edu.ng; +2347036990145

Abstract

Prediction of the market value of a stock is of great interest to investors and policy makers in the stock market [1]. In this study, we focused on precious metal stock prices because some study have shown that precious metals offer market diversification opportunities during period of economic crisis, economic distress and pandemics such as COVID-19 [2]. Previous studies have shown that machine learning techniques can enhance stock price prediction [3] while also time series analyst believed that exogenous variables can improve prediction of stock price [4]. The main question here would be "Do machine learning techniques always outperformed time series approach?" This study predict the stock prices of Wheaton Precious metal using machine learning techniques (Linear regression, Support Vector Regression, Random Forest Regression, Decision Tree Regression and Neural Network regression) and Time series method called ARIMAX (Autoregressive Integrated Moving Average with exogenous variable). To achieve this, daily Wheaton Precious metal stock prices was obtained from a secondary source (www.investing.com) covering 30th December 2004 to 14th March 2023. 80% of the data was used for training the models while 20% of the data set was used for testing. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were used to adjudge the models and the results revealed that machine learning technique (Linear Regression) outperformed all the models with MAE and RMSE values as 0.1726 and 0.2438 respectively while Time series model (ARIMAX) emerged second with MAE and RMSE values as 0.1760 and 0.2454 respectively. Neural Network Regression, Random Forest Regression, Support Vector Regression had MAE and RMSE values as 0.1925 and 0.2663, 0.2153 and 0.3214, and, 0.2353 and 0.2933 respectively. Lastly Decision Tree Regression was the worst model for predicting Wheaton Precious Metal with MAE and RMSE values as 3.7887 and 4.8733 respectively.

[1] P. Soni, Y. Tewari, D. Krishman. J. Physics: Conf. Series. **2161**, 012065 (2022).
[2] H. Alqaralleh, A. Canepa. Res. Policy. 75, **102532** (2022).
[3] M. Vijh, D. Chandola, V. A. Tikkiwal, A. Kumar. Procedia Comp. Sci., **167**:599-606 (2020).
[4] M. O. Adenomon, F. O. Madu. Doi: 10.5772/intechopen.**107979** (2022)

P03

Classification via empirical risk minimisation of power-law distributed clouds

Urte Adomaityte [1], Gabriele Sicuro [1], and Pierpaolo Vivo [1]

[1] *Department of Mathematics, King's College London*

We characterise the learning of a mixture of two clouds of datapoints with generic centroids via empirical risk minimisation (ERM) with any convex loss and regularisation for a large class of non-Gaussian datapoint distributions including power-law distributions. This is done in the setting where the sample size and dimensionality of the dataset are both large, keeping their ratio fixed. We present generalisation and training performance, discuss optimal regularisation and the separability threshold for a choice of fat-tailed data distribution allowing to characterise such phenomenology in the range from data with possibly no covariance to asymptotically recovering the Gaussian data case.

P04

Predicting the effect of training on the internal representations of finite-width Bayesian one-hidden layer networks

In this work, we compute analytically the effect of training on the internal representations of a Bayesian one-hidden layer neural network at finite-width, using the theoretical framework introduced in [1]. In particular, we investigate how the kernel matrix related to these internal representations changes after training. We observe that the kernel matrix elements differ from the infinite-width ones by terms of order $O(N^{-1})$ and we precisely compute this correction. We perform numerical experiments to validate our predictions using Gaussian data, both with random labels and in a teacher-student setting and we present preliminary experiments with regression tasks from the MNIST and CIFAR10 datasets.

References

[1] S Ariosto, R Pacelli, M Pastore, F Ginelli, M Gherardi, and P Rotondo. Statistical mechanics of deep learning beyond the infinite-width limit. *arXiv preprint arXiv:2209.04882*, 2022.

P05

A sunburst of solutions in the continuous negative perceptron

Brandon Livio Annesi[1], **Clarissa Lauditi**[1,2], **Carlo Lucibello**[1], **Enrico M. Malatesta**[1], **Gabriele Perugini**[1], **Fabrizio Pittorino**[1], **Luca Saglietti**[1]

[1]*Artificial Intelligence Lab, Bocconi University, 20136 Milano, Italy*
[2]*Department of Applied Science and Technology, Politecnico di Torino, 10129 Torino, Italy*

The landscape properties of high-dimensional constraint satisfaction problems (CSPs) can completely determine the type of configurations that can be efficiently sampled from their space of solutions. In recent years, empirical studies on the landscape of neural networks have shown that low-lying configurations are often found in complex connected structures, where zero-energy paths between pairs of distant solutions can be constructed. In the present work, we investigate the connectivity of solutions in the negative perceptron, a linear neural network model and a prototype of a non-convex continuous CSP. We introduce a novel analytical method for characterizing the typical energy barriers between groups of configurations sampled from the zero-temperature measure of the problem. We find that, despite the overall non-convexity of the space of solutions, below a critical density of constraints $\alpha_\star$, the geodesic path between any solution and the robust solutions of the problem, located in the interior of the solution space, remains strictly zero-energy. We study the shape and the anisotropy of the connected space of solutions, and numerically characterize a sharp transition where the simple connectivity property breaks down.

P06

Statistical mechanics of deep learning beyond the infinite-width limit

S. Ariosto[1,2], **R. Pacelli**[3], **M. Pastore**[4], **F. Ginelli**[1,2], **M. Gherardi**[5,2], **and P. Rotondo**[5,2]

[1] *Dipartimento di Scienza e Alta Tecnologia and Center for Nonlinear and Complex Systems. Università degli Studi dell'Insubria, Via Valleggio 11, 22100 Como, Italy*
[2] *I.N.F.N. Sezione di Milano, Via Celoria 16, 20133 Milano, Italy*
[3] *Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino, 10129 Torino, Italy*
[4] *Université Paris-Saclay, CNRS, LPTMS, 91405 Orsay, France*
[5] *Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy*

Decades-long literature testifies to the success of statistical mechanics at clarifying fundamental aspects of deep learning. Yet the ultimate goal remains elusive: we lack a complete theoretical framework to predict practically relevant scores, such as the train and test accuracy, from knowledge of the training data. Huge simplifications arise in the infinite-width limit, where the number of units N_ℓ in each hidden layer far exceeds the number P of training examples. This idealisation, however, blatantly departs from the reality of deep learning practice, where training sets are larger than the widths of the networks. In this work, we show one way to overcome these limitations. The partition function for fully-connected architectures, which encodes information about the trained models, can be computed analytically with the toolset of statistical mechanics. The computation holds in the "thermodynamic limit" where both N_ℓ and P are large and their ratio $\alpha_\ell = P/N_\ell$, which vanishes in the infinite-width limit, is now finite and generic. This advance allows us to obtain (i) a closed formula for the generalisation error associated to a regression task in a one-hidden layer network with finite α_ℓ; (ii) an expression of the partition function (technically, via an "effective action") for fully-connected architectures with arbitrary number of hidden layers, in terms of a finite number of degrees of freedom (technically, "order parameters"); (iii) a demonstration that the Gaussian processes arising in the infinite-width limit should be replaced by Student-t processes; (iv) a simple analytical criterion to predict, for a given training set, whether finite-width networks (with ReLU activations) achieve better test accuracy than infinite-width ones. As exemplified by these results, our theory provides a starting point to tackle the problem of generalisation in realistic regimes of deep learning.

[1] S. Ariosto, R. Pacelli, M. Pastore, F. Ginelli, M. Gherardi, P. Rotondo. Preprint arXiv:2209.04882 (2022).

P07

Statistical-Computational Tradeoffs in Mixed Sparse Linear Regression

Gabriel Arpino [1] **and Ramji Venkataramanan** [2]

[1]*University of Cambridge*
[2]*University of Cambridge*

We consider the problem of mixed sparse linear regression with two components, where two k-sparse real signals are to be recovered from n unlabelled noisy linear measurements. The sparsity is allowed to be sublinear in the dimension, and the additive noise is assumed to be independent Gaussian. We establish the existence of an extensive computational barrier for this problem through the method of low-degree polynomials, but show that the problem is computationally hard only in a very narrow symmetric parameter regime. We identify a smooth information-computation tradeoff between the sample complexity n and runtime for any randomized algorithm in this hard regime. Via a simple reduction, this provides novel rigorous evidence for the existence of a computational barrier to solving exact support recovery in sparse phase retrieval with sample complexity n = o(k^2). Our second contribution is to analyze a simple thresholding algorithm which, outside of the narrow regime where the problem is hard, solves the associated mixed regression detection problem in linear time and matches the sample complexity required for (non-mixed) sparse linear regression; this allows the recovery problem to be subsequently solved by state-of-the-art techniques from the dense case. As a special case of our results, we show that this simple algorithm is order-optimal among a large family of algorithms in solving exact signed support recovery in sparse linear regression. To the best of our knowledge, this is the first thorough study of the interplay between mixture symmetry, signal sparsity, and their joint impact on the computational hardness of mixed sparse linear regression.

P08

Employing Deep Learning Model for Malar Region-Based Groupwise Age Ranking

Halleluyah O. Aworinde [1] , Segun Adebayo [1] , Akinwale Akinwunmi [1] and Aderonke B. Sakpere [2]

[1]Bowen University, Iwo, Nigeria
[2] University of Ibadan, Ibadan, Nigeria

Age estimation using facial analysis has enticed huge attention in recent time. However, it is still a challenging task due to various factors including constant changes in the craniofacial regions, the skinny part of the head, biological genes and living habits. To alleviate these challenges, this research proposed an effective GroupWise age ranking technique using the malar region as a biomarker for the age estimation which has been speculated by researchers on this subject matter. In this research, the Flickr-Faces-High Quality (FFHQ) dataset was used in which the faces were grouped into ten classes based on the age estimation performed on the face image dataset. You Only Look Once version 3 (YOLO-V3) algorithm was used to annotate the malar region which the biomarker intended to use for this research. The model was developed using darknet architecture. The model achieved Mean Sum Error MSE loss of 0.12 and 0.18 for the left and right malar regions respectively and Intersection over Union (IoU) of 0.75. The result showed that the left malar region is more efficient for age estimation than the right malar region. The outcome of this research has huge implications as it will enhance various real-life age-based decision-making system.

[1] R. Chikkala, S. Edara, and P. Bhima, "Human Facial Image Age Group Classification Based On Third Order Four Pixel Pattern (TOFP) of Wavelet Image," vol. 16, no. 1, pp. 30–40, 2019.

[2] A. S. Falohun, H. O. Aworinde, A. O. Afolabi, W. O. Ismaila, and O. D. Fenwa, "Fingerprint Phenotyping for Ethnicity Classification: A Generative Deep Learning Perspective," *Science Focus: An International Journal of Biological and Physical Sciences*, vol. 23, no. 1, pp. 84–90, 2018, doi: https://doi.org/10.36293/sfj.2019.0029.

[3] A. S. Al-shannaq and L. A. Elrefaei, "Comprehensive Analysis of the Literature for Age Estimation From Facial Images," *IEEE Access*, pp. 93229–93249, 2019.

[4] M. M. Islam and J. H. Baek, "A Hierarchical Approach toward Prediction of Human Biological Age from Masked Facial Image Leveraging Deep Learning Techniques," *Applied Sciences (Switzerland)*, vol. 12, no. 11, 2022, doi: 10.3390/app12115306.

[5] H. Aworinde and O. F. W. Onifade, "A Soft Computing Model of Soft Biometric Traits for Gender and Ethnicity Classification," *international Journal of Engineering and Manufacturing*, vol. 9, no. 2, pp. 54–63, 2019, doi: 10.5815/ijem.2019.02.05.

[6] H. O. Aworinde, A. O. Afolabi, A. S. Falohun, and O. T. Adedeji, "Performance Evaluation of Feature Extraction Techniques in Multi-Layer Based Fingerprint Ethnicity Recognition System," *Asian Journal of Research in Computer Science*, vol. 3, no. 1, pp. 1–9, 2019, doi: 10.9734/AJRCOS/2019/v3i130084.

[7] J. D. Akinyemi and O. F. W. Onifade, "An ethnic-specific age group ranking approach to facial age estimation using raw pixel features," pp. 1–6, 2017, doi: 10.1109/ths.2016.7819737.

P09

Counter-Adversarial Recall of Synthetic Observations (*CARSO*)

Emanuele Ballarin[1], **Alessio Ansuini**[2], **Luca Bortolussi**[1]

[1] *AILab, Dept. of Mathematics and Geoscience, University of Trieste, Trieste, Italy*
[2] *Data Engineering Laboratory, AREA Science Park, Trieste, Italy*

Vulnerability to *adversarial attacks* [1] constitutes a major hurdle towards ensuring compliance of deep learning systems with modeller-expected behaviour, and their adoption in safety-critical scenarios. In spite of significant theoretical and empirical advances in the field, the problem remains open: with no universal solution, and *adversarial training* [1] as the first-line technique for its mitigation – capable of current *state-of-the-art* results [2].

In this work, inspired by cues from cognitive neuroscience [3], we propose a novel adversarial defence mechanism for supervised image classification models – synergistically complementary to adversarial training – that exploits direct knowledge of the representation of the attacked classifier, and a purpose-assembled *denoising conditional deconvolutional variational autoencoder* architecture, to synthesise samplable reconstructions of the input, tentatively *purified* from adversarial perturbations (if any).

During training, a denoising generative model of the input (provided as batches of balanced *clean* and perturbed images) is learned, conditioned on the (ordered) activation values it produces in an adversarially-pretrained classifier. At inference time, the representation alone is used to condition the process, and the resulting sampled images processed by the very same classifier and finally labelled by majority vote.

Experimental evaluation by a well-established benchmark of varied, strong adaptive attacks [4], with different norm-constraints, and across different image datasets and classifier architectures, shows that method proposed is able to defend the classifier significantly better than *state-of-the-art* adversarial training alone – with a tolerable *clean* accuracy toll, and additional computational costs comparable to adversarial training – even against unforeseen attacks, while also effectively shielding the entire defensive architecture (which can be made differentiable, explicitly or in approximation) from the same end-to-end attacks adapted to fool stochastic defences [5].

[1] I.J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and Harnessing Adversarial Examples"; International Conference on Learning Representations, **3** (2015).
[2] S.A. Rebuffi, S. Gowal, D.A. Calian, F. Stimberg, O. Wiles, T.A. Mann, "Data Augmentation Can Improve Robustness"; Advances in Neural Information Processing Systems, **34** (2021).
[3] J.J. Medina, "The Biology of Recognition Memory"; Psychiatric Times, **6**, pp. 13, 14 (2008).
[4] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, M. Hein, "`RobustBench`: a standardized adversarial robustness benchmark"; NeurIPS 2021, **34**, Datasets and Benchmarks Track (2021).
[5] A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, "Synthesizing Robust Adversarial Examples"; Proceedings of the International Conference on Machine Learning, **35** (2018).

P10

Abstract for poster

L.Bardone[1] **and S.Goldt**[1]

[1] *SISSA*

Title: Beyond Gaussian models: a mathematical framework for the analysis of real-world data structures

Neural Network algorithms are very good at extracting information from real-world data. But what makes them perform so well where other algorithms fail? Recent works have linked the success of Neural Networks to their ability to easily access properties of the data distribution that are more complex than the mean vector and the covariance matrix (the high order cumulants-HOC).

To investigate the role of HOC for inference tasks, we consider a hypothesis testing problem that compares Gaussian noise with a non-Gaussian whitened distribution. The aim is to bound the sample complexity (SC), i.e. the number of samples necessary to understand which of the two distributions data was drawn from. In this non-Gaussian model the first two cumulants of the data give no information, which is in fact concentrated on HOC. Hence random matrix results, usually employed for similar Gaussian models, cannot be readily applied and new ways to get SC bounds need to be found. In the poster, we provide an overview on possible methods to achieve these estimates, including techniques based on the Low-Degree Likelihood Ratio.

P11

Relating Implicit Fourier Bias and Adversarial Attacks through Intrinsic Dimension

Lorenzo Basile[1,2], **Nikos Karantzas**[3], **Alberto D'Onofrio**[1]

Luca Bortolussi[1], **Alex Rodriguez**[1,2,†], **Fabio Anselmi**[1,4,†]

[1] *University of Trieste, Trieste (Italy)*
[2] *The Abdus Salam International Centre for Theoretical Physics, Trieste (Italy)*
[3] *Baylor College of Medicine, Houston TX (USA)*
[4] *Massachusetts Institute of Technology, Cambridge MA (USA)*
[†] *Equal senior authorship*

Despite the enormous success of artificial neural networks (ANNs), their predictions are still very fragile and can be drastically changed by subtle perturbations of their inputs, called adversarial attacks [1]. Interestingly, recent results have shown a deep link between the implicit bias of ANNs and their adversarial robustness [2]. However, besides simple models [3] it is extremely difficult to mathematically characterize the implicit bias of a neural network. The work in [4] proposes an algorithm to study an aspect of the implicit bias for complex networks in terms of the essential input frequencies needed to preserve the performance of a trained network. These frequencies are computed by training, for each input image, a learnable modulatory mask that filters the frequency content of the image, reducing it to the bare minimum required to preserve correct classification. The essential frequency masks constitute a fingerprint of the network, since they encode the information content that the ANN is really "looking" at when processing the input.

In this work we leverage this methodology to investigate the frequency relation between adversarial attacks and implicit bias in complex networks. In particular, we train both essential frequency masks (using a very similar approach to [4]) and modulatory masks capable of *reverting the effect of an adversarial perturbation* in order to recover the correct prediction. We use these two sets of masks to test the hypothesis that the network bias in the Fourier domain is highly correlated with the frequencies used to revert the effect of the adversarial attack and recover the correct class. The final goal is to provide empirical evidence that the network bias determines the nature of the adversarial attacks, in the same spirit of [2].

However, the definition and computation of such correlation pose significant difficulties since the two sets of modulatory masks contain high dimensional objects, and their correlation is expected to be highly non-linear. To overcome such difficulties we introduce a *novel non-linear correlation method* based on intrinsic dimension estimation. Our results show a strong correlation between the two data manifolds, thus empirically demonstrating the link between the network bias in Fourier space and the target frequencies of the adversarial attacks.

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks", arXiv preprint arXiv:1312.6199 (2013).

[2] F. Faghri, S. Gowal, C. Vasconcelos, D. J. Fleet, F. Pedregosa, and N. Le Roux, "Bridging the gap between adversarial robustness and optimization bias", ICLR Workshop on Security and Safety in Machine Learning Systems (2021).

[3] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, "Implicit bias of gradient descent on linear convolutional networks", Advances in Neural Information Processing Systems 31 (2018)

[4] N. Karantzas, E. Besier, J. Ortega Caro, X. Pitkow, A. S. Tolias, A. B. Patel, and F. Anselmi, "Understanding Robustness and Generalization of Artificial Neural Networks Through Fourier Masks", Frontiers in Artificial Intelligence 5 (2022).

P12

Probability flow solution of the Fokker-Planck equation

Nicholas M. Boffi and Eric Vanden-Eijnden

Courant Institute of Mathematical Sciences, New York University

The method of choice for integrating the time-dependent Fokker-Planck equation in high-dimension is to generate samples from the solution via integration of the associated stochastic differential equation. Here, we study an alternative scheme based on integrating an ordinary differential equation that describes the flow of probability. Acting as a transport map, this equation deterministically pushes samples from the initial density onto samples from the solution at any later time. Unlike integration of the stochastic dynamics, the method has the advantage of giving direct access to quantities that are challenging to estimate from trajectories alone, such as the probability current, the density itself, and its entropy. The probability flow equation depends on the gradient of the logarithm of the solution (its "score"), and so is a-priori unknown. To resolve this dependence, we model the score with a deep neural network that is learned on-the-fly by propagating a set of samples according to the instantaneous probability current. We show theoretically that the proposed approach controls the KL divergence from the learned solution to the target, while learning on external samples from the stochastic differential equation does not control either direction of the KL divergence. Empirically, we consider several high-dimensional Fokker-Planck equations from the physics of interacting particle systems. We find that the method accurately matches analytical solutions when they are available as well as moments computed via Monte-Carlo when they are not. Moreover, the method offers compelling predictions for the global entropy production rate that out-perform those obtained from learning on stochastic trajectories, and can effectively capture non-equilibrium steady-state probability currents over long time intervals. We discuss how our theory leads to insights for the development of generative models, as well as demonstrate recent extensions to compute the entropy production rate for a $16,384$-dimensional Fokker-Planck equation arising in the study of motility-induced phase separation in active matter.

[1] N. M. Boffi and Eric Vanden-Eijnden. "Probability flow solution of the Fokker-Planck equation," arXiv:2206.04642.

[2] Michael S. Albergo*, N. M. Boffi*, and Eric Vanden-Eijnden. "Stochastic Interpolants: A Unifying Framework for Flows and Diffusions," arXiv:2303.08797.

[3] N. M. Boffi, and Eric Vanden-Eijnden. "Computing the local entropy production rate of non-equilibrium active matter systems with machine learning," *in preparation*.

P13

Beyond the Universal Law of Robustness: Sharper Laws for Random Features and Neural Tangent Kernels

P14

Thermodynamics of attractor neural networks:
from bi-directional to models with coupled-replicas

Adriano Barra, Giovanni Catania[1], Aurélien Decelle, Beatriz Seoane

[1] *Departamento de Física Teórica, Universidad Complutense de Madrid, 28040 Madrid, Spain.*

The first part of my presentation concerns an exhaustive statistical mechanics treatment of bi-directional associative memories (BAMs), a simple generalization of the Hopfield model to a bipartite structure that allows for the efficient retrieval of pairs of associated patterns (e.g. a person's face and their name). By exploiting rigorous statistical physics techniques based on spin-glass theory, a detailed characterization of the equilibrium phase diagram of the model is discussed, in the thermodynamic limit where the size of both layers and the number of pattern pairs to be embedded grow to infinity at finite ratios. Such analysis generalized the mean-field theory for the Hopfield model developd by Amit, Gutfreund and Sompolinsky. An analytic and numerical analysis of the different transition curves is carried out in terms of the asymmetry between the two layers, both using the replica symmetric (RS) and the 1-step replica symmetry breaking (1-RSB) ansatzs: in particular, the $0-$ temperature critical load is quantified using the two ansatzs. This work [1] represents a first step in considering a learning scheme of Restricted Boltzmann machines (RBMs) with a given structure of the weight matrix in terms of a Hebbian-like decomposition, allowing for the learning of pattern components through gradient ascent on the likelihood function. In the second part, I discuss a general framework to characterize how the equilibrium phase diagram of spin glass models might modified when instead of consider one copy of the original system on its own, a certain number of them is taken with the same quenched disorder, with an additional ferromagnetic coupling between degrees of freedom belonging to different "replicas" of the original model.

References

[1] A. Barra, G. Catania, A. Decelle, B. Seoane, Thermodynamics of bidirectional associative memories, arXiv:2211.09694

P15

Large deviations in stochastic dynamics over graphs through Matrix Product Belief Propagation

Stefano Crotti[1], **Alfredo Braunstein**[1,2,3]

[1] *Politecnico di Torino, Corso duca degli Abruzzi 24, 10124 Torino*
[2]*Italian Institute for Genomic Medicine, IRCCS Candiolo, SP-142, I-10060, Candiolo (TO), Italy*
[3]*INFN, Sezione di Torino, Torino, Italy*

Stochastic processes on graphs can describe a great variety of phenomena ranging from vehicular traffic to neural activity and epidemic spreading. While many existing methods [1, 2, 3, 4] can accurately describe typical realizations of such processes, computing properties of extremely rare events is generally a hard task. This poster introduces the Matrix Product Belief Propagation approximation [5], which can be applied to Markov processes biased by arbitrary reweighting factors that concentrate most of the probability mass on rare events. Our approach builds on a version of the dynamic cavity method [6] where distributions of pairs of trajectories are parametrized by matrix product states, a well-known tool in the field of many-body quantum systems. Some applications are mentioned: Bayesian inference on epidemic models and the analysis of Glauber dynamics of Ising models.

[1] N. Antulov-Fantulin, A. Lancic, T. Smuc, H. Stefancic, and M. Sikic, Physical Review Letters **114**, 24 (2015)
[2] I. Neri and D. Bollé, Journal of Statistical Mechanics: Theory and Experiment **2009**, P08009 (2009).
[3] B. Karrer and M. E. Newman, Physical Review E **016101** (2010).
[4] M. Shrestha, S. V. Scarpino, and C. Moore, Physical Review E **92**, 022821 (2015).
[5] S. Crotti, and A. Braunstein, arXiv:2303.17403 (2023).
[6] T. Barthel, C. De Bacco, and S. Franz, Physical Review E **97**, 010104 (2018).

P16

Learning (with) Deep Random Networks of Extensive Width

Hugo Cui[1], **Florent Krzakala**[2], **Lenka Zdeborová**[1], **Dominik Schröder**[3], **Daniil Dmitriev**[3], **Bruno Loureiro**[4]

[1]*(Presenting author underlined) SPOC lab, EPFL, Switzerland*
[2] *IdePhics lab, EPFL, Switzerland*
[3] *Department of Mathematics, ETH, Switzerland*
[4] *Department of Informatics, ENS, France*

We first consider the problem of learning a target function corresponding to a deep, extensive-width, non-linear neural network with random Gaussian weights [1]. We consider the asymptotic limit where the number of samples, the input dimension and the network width are proportionally large. We derive a closed-form expression for the Bayes-optimal test error, for regression and classification tasks. We contrast these Bayes-optimal errors with the test errors of ridge regression, kernel and random features regression. We find, in particular, that optimally regularized ridge regression, as well as kernel regression, achieve Bayes-optimal performances, while the logistic loss yields a near-optimal test error for classification. We further show numerically that when the number of samples grows faster than the dimension, ridge and kernel methods become suboptimal, while neural networks achieve test error close to zero from quadratically many samples.

We then study the problem of learning the deep random network target using a fully connected network with frozen intermediate layers and trainable readout layer [2]. This problem can be seen as a natural generalization of the widely studied random features model to deeper architectures. First, we prove Gaussian universality of the test error in a ridge regression setting where the learner and target networks share the same intermediate layers, and provide a sharp asymptotic formula for it. Establishing this result requires proving a deterministic equivalent for traces of the deep random features sample covariance matrices which can be of independent interest. Second, we conjecture the asymptotic Gaussian universality of the test error in the more general setting of arbitrary convex losses and generic learner/target architectures. We provide extensive numerical evidence for this conjecture, which requires the derivation of closed-form expressions for the layer-wise post-activation population covariances. In light of our results, we investigate the interplay between architecture design and implicit regularization.

[1] H. Cui, F. Krzakala, L.Zdeborová, arXiv: 2302.00375 (2023).
[2] D. Schröder, H.Cui, D.Dmitriev, B.Loureiro arXiv:2302.00401 (2023).

P17

Robust Inference of Causal Links in High-Dimensional Dynamical Processes from the Information Imbalance of Rank Statistics

Vittorio Del Tatto[1], **Gianfranco Fortunato**[1], **Domenica Bueti**[1], **and Alessandro Laio**[1]

[1] *Scuola Internazionale Superiore di Studi Avanzati, SISSA, via Bonomea 265, 34136 Trieste, Italy*

We introduce an approach which allows inferring causal relationships between variables for which the time evolution is available. The idea at the core of the method is the same of Granger Causality [1], namely testing if including the information of a putative driver system X can improve the predictability of a putative driven system Y. However, the test is built on a statistical quantity derived from distance ranks, using the Information Imbalance measure [2]. The method is nonparametric, it requires no assumptions on the underlying dynamics and it makes causality detection possible even for high-dimensional systems where only few of the variables are known or measured. Benchmark tests on coupled dynamical systems demonstrate that our approach outperforms other model-free causality detection methods, successfully handling both unidirectional and bidirectional couplings. Moreover, we show that the method can be used to robustly detect causality in EEG experiments.

[1] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods", *Econometrica*, **37**, 3 (1969).

[2] A. Glielmo, C. Zeni, B. Cheng, G. Csányi, and A. Laio, "Ranking the information content of distance measures", *PNAS Nexus*, **1**, 04 (2022).

P18

Deterministic equivalent and error universality of deep random features learning

Dominik Schröder[1], **Hugo Cui**[2], **Daniil Dmitriev**[3], **and Bruno Loureiro**[4]

[1]*Department of Mathematics, ETH Zurich*
[2]*Statistical Physics Of Computation lab., Institute of Physics, Ecole Polytechnique Fédérale de Lausanne (EPFL)*
[3]*Department of Mathematics, ETH Zurich and ETH AI Center*
[4]*Departement d'Informatique, École Normale Supérieure (ENS) - PSL CNRS*

This manuscript considers the problem of learning a random Gaussian network function using a fully connected network with frozen intermediate layers and trainable readout layer. is problem can be seen as a natural generalization of the widely studied random features model to deeper architectures. First, we prove Gaussian universality of the test error in a ridge regression seing where the learner and target networks share the same intermediate layers, and provide a sharp asymptotic formula for it. Establishing this result requires proving a deterministic equivalent for traces of the deep random features sample covariance matrices which can be of independent interest. Second, we conjecture the asymptotic Gaussian universality of the test error in the more general seing of arbitrary convex losses and generic learner/target architectures. We provide extensive numerical evidence for this conjecture. In light of our results, we investigate the interplay between architecture design and implicit regularization

P19

Neural-prior stochastic block model

O. Duranthon[1] **and L. Zdeborová**[1]

[1] *SPOC laboratory, EPFL, Lausanne, Switzerland*

The stochastic block model (SBM) is widely studied as a benchmark for graph clustering aka community detection. In practice, graph data often come with node attributes that bear additional information about the communities. Previous works modeled such data by considering that the node attributes are generated from the node community memberships. In [1] the attributes were generated via logistic regression on the community membership, in the contextual-SBM [2, 3] the communities determine centroids for a Gaussian mixture model generating the attributes.

In this work, motivated by a recent surge of works in signal processing using deep neural networks as priors, we propose to model the communities as being determined by the node attributes rather than the opposite. We define the corresponding model; we call it the neural-prior SBM. We propose an algorithm, stemming from statistical physics, based on a combination of belief propagation and approximate message passing. We analyze the performance of the algorithm as well as the Bayes-optimal performance. We identify detectability and exact recovery phase transitions, as well as an algorithmically hard region. The proposed model and algorithm can be used as a benchmark for both theory and algorithms. To illustrate this, we compare the optimal performances to the performance of simple graph neural networks.

[1] J. Yang, J. McAuley and J. Leskovec, IEEE 13th int. conf. on data mining (2013).
[2] N. Binkiewicz, J. T. Vogelstein and K. Rohe, Biometrika **104** (2017).
[3] Y. Deshpande, S. Subhabrata, A. Montanari and E. Mossel, Advances in Neural Inf. Proc. Syst. **31** (2018).

P20

Unified field theoretical approach to deep and recurrent neural networks

Kai Segadlo[1,2,3], **Bastian Epping**[2,4], **Alexander van Meegen**[2,5,6], **David Dahmen**[2], **Michael Krämer**[4], **Moritz Helias**[1,2]

[1] *Department of Physics, Faculty 1, RWTH Aachen University, Aachen, Germany*
[2] *Institute of Neuroscience and Medicine (INM-6) and Institute for Advanced Simulation (IAS-6) and JARA-Institute Brain Structure-Function Relationships (INM-10), Jülich Research Centre, Jülich, Germany*
[3] *Georg-August University Göttingen, Göttingen, Germany*
[4] *Institute for Theoretical Particle Physics and Cosmology, RWTH Aachen University, Aachen, Germany*
[5] *Center for Brain Science, Havard University, Cambridge, Massachusetts*
[6] *Institute of Zoology, University of Cologne, Cologne, Germany*

The recent progress in the field of neural networks and machine learning in general, due to deep learning, has set new standards across a broad range of domains from image processing to speech recognition [1]. However, much of the up-to-date research is of empirical nature and hence raises questions about guarantees for accuracy and robustness. Bayesian inference on Gaussian processes [2] has proven to be a viable approach for studying deep neural networks (DNNs) in the infinite-width limit [3]. While the equivalence of DNNs and Gaussian processes is known and commonly proven with the central limit theorem, addressing recurrent neural networks (RNNs) is more involved due to weight sharing across timesteps.

To address the weight distribution problem, we start from first principles and develop a field-theoretic formalism that allows us to study both DNNs and RNNs in a Bayesian framework. We draw on well established methods from statistical physics of disordered systems [4]; in particular, we employ a saddle point approximation, which becomes exact in the infinite-width limit ($n \to \infty$), to calculate the prior distribution of network outputs. This results in mean-field equations that determine the Gaussian process kernels for the two network architectures.

The kernels differ in general but are equivalent when considering activations at equal time points or layers, respectively. Intriguingly, this equivalence implies equivalent performance of Bayesian inference in DNNs and RNNs in the $n \to \infty$ limit if readouts are taken at a fixed time point or layer, respectively. However, the mean-field equations for the two architectures differ in their temporal structure: The RNN shows temporal correlations due to weight sharing, whereas the activations of the DNN are uncorrelated across different layers.

The new formalism allows us to systematically compute finite-size corrections globally across all layers or time steps for any activation function in powers of $1/n$. The leading order corrections are found to be different for the two architectures, even at equal layers or times.

A manuscript of this work is published in [5].

[1] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature*. **521**, 436-444 (2015,5), https://doi.org/10.1038/nature14539
[2] Rasmussen, C. & Williams, C. Gaussian processes for machine learning.. (MIT Press,2006)
[3] Lee, J., Bahri, Y., Novak, R., Schoenholz, S., Pennington, J. & Sohl-Dickstein, J. Deep Neural Networks as Gaussian Processes. *International Conference On Learning Representations*. (2018)
[4] Hertz, J., Roudi, Y. & Sollich, P. Path integral methods for the dynamics of stochastic and disordered systems. *Journal Of Physics A: Mathematical And Theoretical*. **50**, 033001 (2016,12), https://dx.doi.org/10.1088/1751-8121/50/3/033001
[5] Segadlo, K., Epping, B., Meegen, A., Dahmen, D., Krämer, M. & Helias, M. Unified field theoretical approach to deep and recurrent neuronal networks. *Journal Of Statistical Mechanics: Theory And Experiment*. **2022**, 103401 (2022,10), https://dx.doi.org/10.1088/1742-5468/ac8e57

P21

Statistical mechanics of the maximum-average submatrix problem

Vittorio Erba[1], **Florent Krzakala**[2], **Rodrigo Pérez**[2] **and Lenka Zdeborová**[1]

[1]*Statistical Physics of Computation Laboratory,*
[2]*Information, Learning and Physics Laboratory,*
École polytechnique fédérale de Lausanne (EPFL) CH-1015 Lausanne

We study the maximum-average submatrix problem, in which given an $N \times N$ matrix J one needs to find the $k \times k$ submatrix with the largest average of entries. We study the problem for random matrices J whose entries are i.i.d. random variables by mapping it to a variant of the Sherrington-Kirkpatrick spin-glass model at fixed magnetization. We characterize analytically the phase diagram of the model as a function of the submatrix average and the size of the submatrix k in the limit $N \to \infty$. We consider submatrices of size $k = mN$ with $0 < m < 1$. We find a rich phase diagram, including dynamical, static one-step replica symmetry breaking and full-step replica symmetry breaking. In the limit of $m \to 0$, we find a simpler phase diagram featuring a frozen 1-RSB phase, where the Gibbs measure is composed of exponentially many pure states each with zero entropy.

[1] Erba, V., Krzakala, F., Pérez, R., and Zdeborová, L. (2023). Statistical mechanics of the maximum-average submatrix problem. arXiv preprint arXiv:2303.05237.

P22

(S)GD over Diagonal Linear Networks: Implicit Regularisation, Large Stepsizes and Edge of Stability

Mathieu Even[1], **Scott Pesme**[2], **Suriya Gunasekar**[3] **and Nicolas Flammarion**[2]

[1] *(Presenting author underlined) ENS and Inria Paris*
[2] *EPFL*
[3] *Microsoft Research*

In this paper, we investigate the impact of stochasticity and large stepsizes on the implicit regularisation of gradient descent (GD) and stochastic gradient descent (SGD) over diagonal linear networks. We prove the convergence of GD and SGD with macroscopic stepsizes in an overparametrised regression setting and characterise their solutions through an implicit regularisation problem. Our crisp characterisation leads to qualitative insights about the impact of stochasticity and stepsizes on the recovered solution. Specifically, we show that large stepsizes consistently benefit SGD for sparse regression problems, while they can hinder the recovery of sparse solutions for GD. These effects are magnified for stepsizes in a tight window just below the divergence threshold, in the "edge of stability" regime. Our findings are supported by experimental results.

P23

DFA for Continual Learning

Sara Folchini[1], **Sebastian Goldt**[1]

[1] *SISSA, Trieste*

Continual learning (CL) , or lifelong learning, considers learning through a sequence of tasks, where the learning method has to retain knowledge about past tasks and leverage on it.

In the multi-task scenario it was shown that architectural, regularization and rehearsal strategies can be used to train deep models sequentially on a number of isolated disjoint tasks without forgetting previously acquired knowledge.

However, these strategies are still unsatisfactory if the tasks are not disjoint but constitute a single incremental task (e.g., class-incremental learning).

An emergent and biologically plausible algorithm, Direct Feedback Alignment (DFA), is proposed in the context of Continual Learning. DFA propagates the error through fixed random feedback connections directly to the hidden layers.

Its characteristic local updates are believed to govern synaptic weight updates in the brain and different backward and forward matrices resemble synaptic asymmetry.

When tested on multi-task and class-incremental benchmark datasets, it outperformed existing regularization strategies.

This is achieved combining an architectural strategy with the fact that the random feedback matrix has a strong influence on the direction of the training trajectory and acts as an implicit regularization strategy. The feedback matrix can be kept constant throughout the CL stages, or it can be reinitialized to an orthogonal state for every task switch. We show that this flexibility is the key for the method to perform in the class-incremental scenario.

P24

Understanding the effects of class imbalance on the dynamics of supervised learning

Emanuele Francazi[1], **Marco Baity-Jesi**[2], **Aurelien Lucchi**[3]

[1] *Physics Department, EPFL*
[2] *SIAM Department, Eawag (ETH)*
[3] *Department of Mathematics and Computer Science, University of Basel*

Data imbalance is a common problem in machine learning [1, 2, 3, 4] that can have a critical effect on the performance of a model. Various solutions exist but their impact on the learning dynamics is not understood [5, 6, 7]. We will elucidate the significant negative impact of data imbalance on learning, showing that the learning curves for minority and majority classes follow sub-optimal trajectories when training with a gradient-based optimizer. This slowdown is related to the imbalance ratio and can be traced back to a competition between the optimization of different classes. We will delve into the effects of class imbalance on learning dynamics by seeing how these effects are fundamentally different in gradient descent (GD) and stochastic gradient descent (SGD). Through this analysis, we will not only be able to understand the potential and limitations of some new variants of (S)GD but also the reason for the effectiveness of previously used solutions for class imbalance such as oversampling.

[1] Van Horn, Grant, and Pietro Perona. "The devil is in the tails: Fine-grained classification in the wild." arXiv preprint arXiv:1709.01450 (2017).
[2] Liu, Shigang, et al. "Addressing the class imbalance problem in twitter spam detection using ensemble learning." Computers & Security 69 (2017): 35-49.
[3] Makki, Sara, et al. "An experimental study with imbalanced classification approaches for credit card fraud detection." IEEE Access 7 (2019): 93010-93022.
[4] D'souza, Daniel, et al. "A tale of two long tails." arXiv preprint arXiv:2107.13098 (2021).
[5] Johnson, Justin M., and Taghi M. Khoshgoftaar. "Survey on deep learning with class imbalance." Journal of Big Data 6.1 (2019): 1-54.
[6] Chawla, Nitesh V., Nathalie Japkowicz, and Aleksander Kotcz. "Special issue on learning from imbalanced data sets." ACM SIGKDD explorations newsletter 6.1 (2004): 1-6.
[7] Sagawa, Shiori, et al. "An investigation of why overparameterization exacerbates spurious correlations." International Conference on Machine Learning. PMLR, 2020.

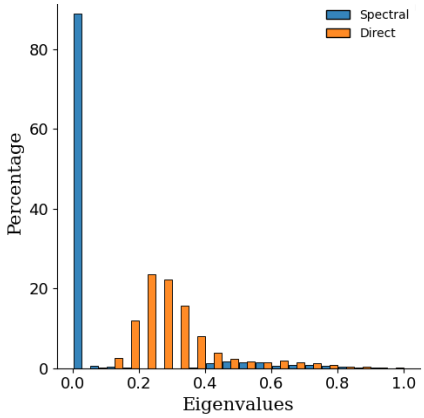
P25

Spectral L2 Regularization of Feedforward Neural Networks

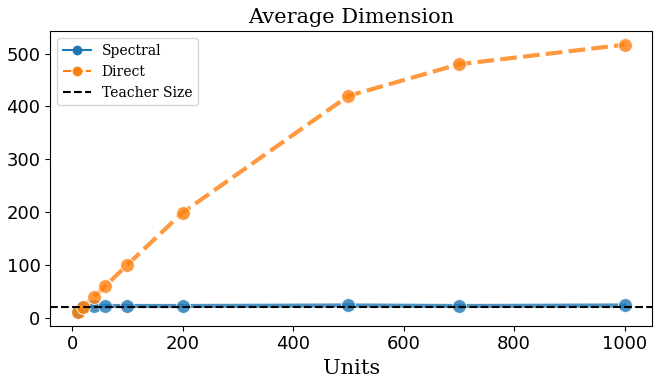
Lorenzo Giambagli[1,2], **Lorenzo Buffoni**[1], **Lorenzo Chicchi**[1], **and Duccio Fanelli**[1]

[1]*CSDC, INFN, Department of Physics and Astronomy, University of Florence, Italy*
[2]*naXys, Namur Center for Complex Systems, University of Namur, Belgium*

Deep Feedforward Neural Networks (FFNNs) are a central tool in the field of Machine Learning. They are typically trained in node space, where the weights are adjusted through suitable optimization protocols. Recently a new approach has been proposed [1], where the learning process is anchored to reciprocal space, and the optimization process targets eigenvectors and eigenvalues of every transfer operator between layers, i.e., the network eigenmodes. In this work, we highlight the efficient spectral parametrization of a Dense layer, which incurs no significant additional computational cost. By focusing on these fundamental mathematical structures a correspondence between eigenvalues and nodes has been made, enabling a better understanding of their pivotal role in training sparse and small FFNNs [2, 3]. Specifically, we demonstrate how adding an $L2$ regularization term that acts on the eigenvalues, next to the classic weight decay, alters the learning dynamics. Indeed, small and compact solutions, rather than large sparse ones, emerges from the learning process without affecting performances. Our results are validated in the teacher-student framework where two hidden layer teacher and student network are used. The second hidden layer size is left constant and trained conventionally, whereas the first is trained either in the Spectral or classical way. Analysing the histogram of the regularized eigenvalues after training reveals a peculiar distribution (see Figure 1a) with a large amount of zeros when compared to an equivalent indicator for the direct space (the feature norm). Removing the nodes in correspondence with such zero-valued eigenvalues not only does not affect the performance of the FFNN but also reveals the central structure of the NN which, indeed is of the same size as the teacher. Such propriety is kept even for layers orders of magnitude bigger than the proper one (see figure 1b). We also show how topological proprieties of the network can be more easily accessed inspecting the trained eigenvectors.



(a) Eigenvalues histogram



(b) Average size of first hidden layer after node removal across 30 trials, variance is negligible.

[1] Machine learning in spectral domain *L. Giambagli, L. Buffoni, T. Carletti, W. Nocentini and D. Fanelli*, Nature Communications (2021).

[2] Spectral pruning of fully connected layers *L. Buffoni, E. Civitelli, L. Giambagli, L. Chicchi and D. Fanelli*, Scientific Reports (2022).

[3] Training of sparse and dense deep neural networks: Fewer parameters, same performance, *L. Chicchi, L. Giambagli, L. Buffoni, T. Carletti, M. Ciavarella, and D. Fanelli* PRE (2021).

P26

Assisted Learning of Energy-Based Models with Normalizing Flows

Louis Grenioux[1], **Éric Moulines**[1], **Marylou Gabrié**[1]

[1]*CMAP, École polytechnique*

Energy-based models (EBM) are versatile unnormalized density estimation models. EBMs can be estimated using maximum likelihood but this procedure inherently requires sampling the learned model. Meanwhile, the energy function is often parametrized by a deep neural network so as to model complicated target distributions, which ultimately defines a high-dimensional sampling problem inheriting the multimodality of the target. As a result maximum likelihood training of EBMs remains a challenge for common MCMC samplers. In this work, we propose to use recent algorithms from the adaptive MCMC literature by jointly training a normalizing flow to enhance the sampling of the intermediate energy function. Unlike commonly used Langevin samplers, adaptive samplers combine local transitions and global jumps in the Markov chain thus enabling much better mixing in multi-modal contexts. Moreover, this auxiliary generative model provides an easy way to get diverse samples realistically reflecting the multi-modality of the energy-based model. We show that the trained models are capable of sampling and modeling complex distributions in high-dimensions.

[1] Erik Nijkamp, Ruiqi Gao, Pavel Sountsov, Srinivas Vasudevan, Bo Pang, Song-Chun Zhu, and Ying Nian Wu. MCMC should mix: Learning energy-based model with neural transport latent space MCMC. *In International Conference on Learning Representations, 2022*.

[2] Jianwen Xie, Yaxuan Zhu, Jun Li, and Ping Li. A tale of two flows: Cooperative learning of langevin flow and normalizing flow toward energy-based model. *In International Conference on Learning Representations, 2022*.

[3] Louis Grenioux, Alain Durmus, Éric Moulines, and Marylou Gabrié. On sampling with approximate transport maps, 2023 *arXiv:2302.04763*.

[4] Matthew D Hoffman, Pavel Sountsov, Joshua V. Dillon, Ian Langmore, Dustin Tran, and Srinivas Vasudevan. NeuTra-lizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport. *In 1st Symposium on Advances in Approximate Bayesian Inference, 2018 1–5, 2019*.

[5] Marylou Gabrié, Grant M. Rotskoff, and Eric Vanden-Eijnden. Adaptive monte carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences, 119(10):e2109420119, 2022*.

P27

Interplay of two finite reservoirs in a bidirectional transport system

Ankita Gupta[1], Bipasha Pal[2], Arvind Kumar Gupta[3]
[1,3]Department of Mathematics, Indian Institute of Technology Ropar, India-140001
[2]Department of Mathematics, Stockholm University, Stockholm 10691, Sweden
[1]2018maz0006@iitrpr.ac.in, [2]bipasha.pal@math.su.se, [3]akgupta@iitrpr.ac.in

Over decades, there has been a great deal of interest in stochastic transport phenomena of various complex systems such as intracellular transport of cargo vesicles and vehicular flow, both theoretically and physically. Particles either self-driven or driven by some external field traveling stochastically along a one or multi-dimensional lattice have been utilized to model various transport processes, both natural as well as man-made. In order to analyze the propelled dynamics of these driven diffusive systems, Totally Asymmetric Simple Exclusion Process (TASEP) [1,2] is widely used as the most prominent paradigm of the driven models to examine several stationary system features.

Motivated by the interplay of multiple species in several real world transport processes, we propose a bidirectional totally asymmetric simple exclusion process with two finite particle reservoirs regulating the inflow of oppositely directed particles corresponding to two different species. The system's stationary characteristics such as densities, currents, etc., are investigated using a theoretical framework based on mean-field approximation and are supported by extensive Monte Carlo simulations. The impact of individual species populations, quantified by filling factor, has been comprehensively analyzed considering both equal and unequal conditions. For the equal case, the system exhibits the spontaneous symmetry-breaking phenomena and admits both symmetric as well as asymmetric phases. Moreover, the phase diagram exhibits a new asymmetric phase and displays a non-monotonic variation in the number of phases with respect to the filling factor. For unequal filling factors, the phase schema can display at most five phases including a new phase that shows maximal current for one of the species.

References

[1] B. Derrida, *An exactly soluble non-equilibrium system: the asymmetric simple exclusion process*, Phys. Rep. 301, 65, 1998.

[2] A. Schadschneider, D. Chowdhury and K. Nishinari *Stochastic transport in complex systems: from molecules to vehicles*, Elsevier, 2010.

P28

Algorithmic Threshold for Multi-Species Spherical Spin Glasses

Brice Huang
Massachusetts Institute of Technology

Abstract

We study efficient optimization of the Hamiltonians of multi-species spherical spin glasses. Our results characterize the maximum value attained by algorithms that are suitably Lipschitz with respect to the disorder through a variational principle that we study in detail. We rely on the branching overlap gap property introduced in our previous work and develop a new method to establish it that does not require the interpolation method. Consequently our results apply even for models with non-convex covariance, where the Parisi formula for the true ground state remains open. As a special case, we obtain the algorithmic threshold for all single-species spherical spin glasses, which was previously known only for even models. We also obtain closed-form formulas for pure models which coincide with the E_∞ value previously determined by the Kac-Rice formula. This is joint work with Mark Sellke (Harvard).

P29

How to count in hierarchical landscapes: quenched complexity for the spherical models

Jaron Kent-Dobias[1], **and Jorge Kurchan**[2]

[1]DYNSYSMATH, *Istituto Nazionale di Fisica Nucleare, La Sapienza, Rome*
[2]*Centre National de la Researche Scientifique, Laboratoire de Physique de l'ENS, Paris*

Complex landscapes are characterized by their many saddle points. Determining their number and organization is a long-standing problem, in particular for tractable Gaussian mean-field potentials, which include glass and spin glass models. The annealed approximation is well understood, but is generically not exact. Here we describe the exact quenched solution for the general case, which incorporates Parisi's solution for the ground state, as it should. The quenched solution also correctly uncovers the full distribution of saddles at a given energy, a structure that is lost in the annealed approximation. This structure should be a guide for the accurate identification of the relevant activated processes in relaxational or driven dynamics.

[1] J Kent-Dobias and J Kurchan, "How to count in hierarchical landscapes: a 'full' solution to mean-field complexity," 2022, arXiv:2207.06161 [cond-mat.stat-mech]

P30

Fundamental Limits of Two-layer Autoencoders, and Achieving Them with Gradient Methods

Alexander Shevchenko[1,*], **Kevin Kögler**[1,*], **Hammed Hassani**[2] **and Marco Mondelli**[1]

[1] *Institute of Science and Technology Austria*
[2] *Department of Electrical and Systems Engineering, University of Pennsylvania*
[*] *Authors contributed equally*

Autoencoders are a popular model in many branches of machine learning and lossy data compression. However, their fundamental limits, the performance of gradient methods and the features learnt during optimization remain poorly understood, even in the two-layer setting. In fact, earlier work has considered either linear autoencoders or specific training regimes (leading to vanishing or diverging compression rates). Our paper addresses this gap by focusing on non-linear two-layer autoencoders trained in the challenging proportional regime in which the input dimension scales linearly with the size of the representation. Our results characterize the minimizers of the population risk, and show that such minimizers are achieved by gradient methods; their structure is also unveiled, thus leading to a concise description of the features obtained via training. For the special case of a sign activation function, our analysis establishes the fundamental limits for the lossy compression of Gaussian sources via (shallow) autoencoders. Finally, while the results are proved for Gaussian data, numerical simulations on standard datasets display the universality of the theoretical predictions.

P31

Impacts of Central-Pacific El Niño and physical drivers on Eastern Pacific tuna using Explainable Artificial Intelligence

Peng Lian[1,2]

[1]*Key Laboratory of Ocean Circulation and Waves, Institute of Oceanology, Chinese Academy of Sciences;*
[2]*University of Chinese Academy of Sciences;*

Correspondence: fisheries@foxmail.com

Abstract:
Bigeye tuna (Thunnus obesus) is a crucial migratory species that forages deeply, and El Niño events highly influence its distribution in the eastern Pacific Ocean. While sea surface temperature (SST) is widely recognized as the main factor affecting bigeye tuna (BET) distribution during El Niño events, the roles of different types of El Niño and subsurface oceanic signals, such as ocean heat content (OHC) and mixed layer depth (MLD), remain unclear. We conducted a spatial-temporal analysis to investigate the relationship among BET distribution, El Niño events, and the underlying oceanic signals to address this knowledge gap. We used monthly purse seine (PS) fisheries data of BET in the Eastern Tropical Pacific Ocean (ETPO) from 1994 to 2012 and extracted Central-Pacific El Niño (CP) indices based on Niño 3 and Niño 4 indexes. Furthermore, we employed Explainable Artificial Intelligence (XAI) models to identify the main patterns and feature importance of the six environmental variables and used information flow analysis to determine the causality between the selected factors and BET distribution. Finally, we analyzed Argo datasets to calculate the vertical, horizontal, and zonal mean temperature differences during CP and normal years to clarify the oceanic thermodynamic structure differences between the two types of years. Our findings reveal that BET distribution during CP years is mainly driven by advection feedback of subsurface warmer thermal signals and vertically warmer habitats in the CP domain area, especially in high-yield fishing areas. The high frequency of CP-type El Niño events will likely lead to the westward shift of fisheries centers.

Keywords: Bigeye tuna, Central-Pacific El Niño, ocean heat content, Argo, XAI, information flow

P32

Statistical Inference for Multivariate Bilinear Panel Data

A. Lmakri[1] **and A. Akharif**[2]

[1]*Hassan II University, ENSAM of Casablanca, Morocco*
[2]*Mathematics and Applications Laboratory, FST, Abdelmalek Essaadi University, Tangier, Morocco*

This poster aims to propose both parametric and nonparametric tests that are locally and asymptotically optimal (in the Hàjek and Le Cam senses) for detecting dependence in multivariate bilinear panel data models. To achieve this, we establish the local asymptotic normality (LAN) property to construct tests that are both locally and asymptotically optimal. Asymptotic relative efficiencies are also computed, and simulations are conducted to examine the performance of these tests in small sample sizes.

[1] A. Akharif, and M. Hallin. Efficient detection of random coefficients in autoregressive models, Ann. Stat., **31**, (2003).
[2] A. Lmakri, A. Akharif, A. Mellouk, and M. Fihri. Pseudo Gaussian and rank based tests for first-order superdiagonal bilinear models in panel data, Revstat Stat. J., **19**, (2021).
[3] L.M. Le Cam. *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York, (1986).
[4] J. Hàjek, and Z. Šidàk. *Theory of Rank Tests*, Academic Press, New York, (1967).

P33

Large population models: scalable generative modeling from multiple non-representative databases

João Loula[1], **Cameron Freer**[1], **Ulrich Schaechtle**[2]„ **Zane Shelby**[2]„ **and Vikash Mansinghka**[1]

[1]*(Presenting author underlined) MIT*
[2]*DG Architecture Lab*

We introduce a nonparametric bayesian framework for modeling large populations sampled by multiple non-representative data sources, naturally handling datasets with mismatched or differently coded variables. We show how inference can be scaled to large databases through a novel sequential Monte Carlo algorithm with GPU acceleration that exploits parallelism over rows and columns. The resulting generative models are symbolic and interpretable, and support conditioning and information-theoretic queries with mixed data types. Using real world patient data combined from different health records, we show empirically that our approach can generate faithful synthetic data, make accurate predictions about small subpopulations, and generalize inferences from non-representative samples.

P34

The exponential capacity of modern associative memories

The Hopfield model is a paradigmatic model of associative memory which is able able to retrieve the stored patterns from noisy observations.
Thanks to the replica method from spin glass theory, it has been shown that the model is able to store a number of uncorrelated patterns that scales linearly with the size of the system, with the asymptotic threshold that can be computed to high precision.

Here we present the statistical physics analysis of a recently proposed generalization of the Hopfield model, named modern Hopfield network (MHN).
The MHN has an exponential capacity, i.e. it is able to store $P = \exp(\alpha N)$ patterns where N is the size of the system, for an exponential rate α small enough.
Besides this huge storage capacity, the MHN is linked to the attention mechanism of Transformer architectures in deep learning. In fact, one step in the dynamics of a MHN can be mapped into the forward pass of an attention layer.

We provide the phase diagram of the model thanks to a large deviation analysis of a Random Energy Model (REM) related to the problem. Our framework allows the analysis of a large class of pattern ensembles, each one inducing a characteristic distribution in the related REM problem. We derive exact thresholds for single pattern retrieval and lower bounds for all patterns retrieval largely improving the existing ones.
Additionally, we are able to compute the exact size of the basins of attraction, and to analyze different scaling regimes of the number of patterns and the coupling strength with N.

The generic framework we present is then applied to the cases of gaussian and spherical patterns. For spherical patterns, we find that the lower bound for the full retrieval is sharp, and that one can arbitrarily increase the capacity by increasing the interaction strength. For Gaussian patterns instead we find a richer picture where the the lower bound and the single pattern retrieval threshold do not match and for any interaction strength there is a limit capacity above which the system enters either a "liquid" phase or a "condensed" spin glass phase.

P35

Evolution in time of the dynamical trajectories of discrete spin systems

D. Machado[1], **E. Dominguez**[2], **E. Aurell**[3], **and R. Mulet**[1]

[1] *Group of Complex Systems and Statistical Physics. Department of Theoretical Physics, Physics Faculty, University of Havana, Cuba*

[2] *Donders Institute for Brain, Cognition and Behaviour. Radboud University, Nijmegen, The Netherlands*

[3] *KTH – Royal Institute of Technology, AlbaNova University Center, SE-106 91 Stockholm, Sweden*

We consider classical spin systems evolving in continuous time with interactions given by a locally tree-like graph. In the recent past, several efforts have been made to describe the system's dynamics by introducing closures to the Master Equation [1, 2, 3, 4]. So far, this approach has given accurate results in traditional models like the Ising ferromagnet on a random graph, but fails to predict the dynamics of models with disordered interactions. In our latest developments of the technique, presented on this poster, we describe the system's behaviour with equations that explicitly consider a non-trivial contribution of all spins' past histories. Each site's dynamics is parameterized using two numbers: the final value of the corresponding spin and its average over time. Then, we derive Master Equations for the time evolution of the probability density of having a given trajectory up to time t. To test it, we explored some simple models: a single Ising spin, a chain of Ising spins, and the ferromagnetic Ising model on a random regular graph. In future collaborations, we aim to apply the new equations to more complicated problems that exhibit disorder.

[1] E. Aurell, G. Del Ferraro, E. Dominguez, and R. Mulet. Phys. Rev. E, 95:052119, (2017)
[2] E. Aurell, E. Dominguez, D. Machado, and R. Mulet. Phys. Rev. Letters, 123:230602, (2019).
[3] D. Machado and R. Mulet. Phys. Rev. E, 104:054303, (2021)
[4] E. Aurell, D. Machado, and R. Mulet. J. Phys. A: Math. Theor, (2023), accepted manuscript, 10.1088/1751-8121/acc8a4

P36

Nonparametric prediction and supervised classification for spatial dependent functional data under fixed sampling design.

P37

Attacks on Online Learners: a Teacher-Student Analysis

Riccardo G. Margiotta [1], **Sebastian Goldt** [1], **and Guido Sanguinetti** [1]

[1] *Scuola Internazionale Superiore di Studi Avanzati, 34136 Trieste, Italy*

Machine learning models are famously vulnerable to adversarial attacks: small ad-hoc perturbations of the data that can catastrophically alter the model predictions. While a large literature has studied the case of test-time attacks on pre-trained models, the important case of attacks in an online learning setting has received little attention so far. In this work, we study the scenario where an attacker can perturb data labels to manipulate the learning dynamics of an online learner, with the aim of biasing the learner towards a nefarious target.

We introduce a teacher-student-target model for this purpose, where the canonical teacher-student setup [1, 2] is enriched by an additional ingredient: the attacker's *target function*. Our focus is on the attacker's perspective, whose purpose is to find the best sequence of perturbations that will bring the student as close as possible to the target. This can be formalised as an optimal control problem [3, 4], where for each training step the attacker tries to minimise two opposing costs: the magnitude of the next perturbation, which crucially depends on the *attack strength*, and the current distance between the student and the target.

We investigate the steady state of the student for different attack strategies, obtaining explicit results for simple linear learners, and exploring numerically the behaviour of more complex learners on real datasets. Our findings show that greedy attacks can be extremely efficient, especially when data stream in small batches. Moreover, we observe a phase transition-like phenomenon occurring when a critical threshold in the attack strength is passed, leading to a collapse in the accuracy of the learner.

[1] H. S. Seung, H. Sompolinsky, N. Tishby, Phys. Rev. A (1992).
[2] S. Lee, S. Goldt, A. Saxe, ICML (2021).
[3] X. Zhang, X. Zhu, L. Lessard, L4DC (2020).
[4] Y. Wang, K. Chaudhuri, ArXiv:1808.08994 (2018).

P38

Inference in conditioned dynamics through causality restoration

Alfredo Braunstein[1,2,3], **Giovanni Catania**[4], **Luca Dall'Asta**[1,2,3,5], **Matteo Mariani**[1,*], **Anna Paola Muntoni**[1]

[1]*DISAT, Politecnico di Torino, Corso Duca Degli Abruzzi 24, 10129 Torino*
[2]*INFN, Sezione di Torino, Torino, Italy*
[3]*Italian Institute for Genomic Medicine, IRCCS Candiolo, SP-142, I-10060 Candiolo (TO), Italy*
[4]*Departamento de Física Téorica I, Universidad Complutense, 28040 Madrid, Spain*
[5] *Collegio Carlo Alberto, P.za Arbarello 8, 10122, Torino, Italy*
[*] matteo.mariani@polito.it

We consider the problem of reconstructing (or inferring) a realization x^* of a high-dimensional stochastic process given some partial observations $\mathcal{O}$ on it. Even when sampling from the stochastic dynamics is feasible, inference is typically computationally hard [1]. This is because the observations restrict in a non trivial way the space of possible instances of the stochastic process. In fact, a sample x of the original stochastic dynamics might be rejected due to inconsistency with the observations $\mathcal{O}$. This ultimately renders the sampling of the stochastic dynamics conditioned to $\mathcal{O}$ non-trivial and inefficient. From a deeper point of view, the observations lead to causality breaking of the sampling process by imposing constraints to the trajectories at all times.
The Causal Variational Approach [2] is proposed as an approximate method to generate without rejection independent samples from the conditioned distribution. The procedure relies on learning the parameters of a generalized dynamical model that optimally describes the conditioned distribution in a precise variational sense. The method allows to efficiently compute observables from the conditioned dynamics by averaging over independent samples and it provides an effective unconditioned distribution that is easy to interpret. This approximation can be applied virtually to any dynamics. An example studied is epidemic inference, which can help to develop effective automatic contact tracing methods [3] to mitigate an outbreak. The results of direct comparison with state-of-the-art inference methods, including the Monte Carlo approaches and mean-field methods [4], are promising.

[1] D. J. MacKay. Information theory, inference and learning algorithms. Cambridge univer- sity press, 2003.
[2] A. Braunstein, G. Catania, L. Dall'Asta, M. Mariani, and A. P. Muntoni. Inference in conditioned dynamics through causality restoration, 2023.
[3] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dorner, M. Parker, D. Bonsall, and C. Fraser. Quantifying SARS-CoV-2 transmission suggests epidemic con- trol with digital contact tracing. Science, 368(6491):eabb6936, May 2020. Publisher: American Association for the Advancement of Science.
[4] A. Baker, I. Biazzo, A. Braunstein, G. Catania, L. Dall'Asta, A. Ingrosso, F. Krzakala, F. Mazza, M. Mézard, A. P. Muntoni, M. Refinetti, S. Sarao Mannelli, and L. Zdeborová. Epidemic mitigation by statistical inference from contact tracing data. Proceedings of the National Academy of Sciences, 118(32):e2106548118, Aug. 2021. Publisher: Proceedings of the National Academy of Sciences.

P39

Exploring the weight space of a perceptron *via* enhanced sampling techniques

M. Mele [1, 2], Roberto Menichetti [1,2], Alessandro Ingrosso[3], and Raffaello Potestio[1,2]

[1] Physics Department, University of Trento, via Sommarive, 14 I-38123 Trento, Italy
[2] INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, I-38123 Trento, Italy
[3] The Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, Italy

The steadily growing computing power has made vast amounts of high-throughput data available, opening new windows into complex systems such as cells, the brain, and human societies. However, while the staggering success of Artificial Intelligence and Machine Learning revealed the potential of neural networks, it also raised crucial theoretical questions about them, first and foremost: how does learning take place? In artificial neural networks, learning translates in tuning a large number of connection weights to minimise a loss function. The assumption of Gaussian i.i.d. inputs has been fundamental to the theoretical and computational study of high-dimensional learning; answering the question of whether and how far this hypothesis is restrictive has now become imperative. In our work we have taken advantage of enhanced sampling methods [1], developed in soft matter physics, to exhaustively explore the loss profile of networks with discrete weights, where the optimization landscape is severely rugged even for simple architectures. These tools have proven to be very powerful as they can be directly applied to real datasets, thus allowing us to explore the impact of dimensionality and structure of the data. In particular, we have investigated 4 widely used benchmark datasets: MNIST, FashionMNIST, CIFAR10 and MNIST1D. For each of them we analysed the role of the structure as well as the impact of the input-output correlation. Additionally, this approach enables us to prove whether or not the universality of linear classification with random labels [2] may be extended also in case of non-convex loss and for energetic states higher than the ground state.

[1] R. Menichetti, M. Giulini, & R. Potestio. Eur. Phys. J. B **94**, 204 (2021)
[2] F. Gerace, F. Krzakala, B. Loureiro, L. Stephan, & L. Zdeborová. arXiv:**2205.**13303 (2022)

P40

Evaluating the risk of extinction of fish populations in natural habitat using Reversible-Jump Markov chain Monte Carlo method: A case study on the decline of Notopterus Chitala in India

Dipali Vasudev Mestry[a]*, Md Aktar Ul Karim[a], Joyita Mukherjee[b], Amiya Ranjan Bhowmick[a]

[a]Department of Mathematics, Institute of Chemical Technology, Mumbai
[b]Krishna Chandra College, Hetampur, Birbhum, West Bengal, India
*Presenting Author

The fish species, *Notopterus Chitala*, has a wide distribution in African and Asian countries. This species has been categorized as endangered (EN) in the Conservation Assessment and Management Plan. The objective of the study is to investigate the cause of the decline of the species in their natural habitat. A recent survey conducted by Conservation Assessment and Management Plan (NBFGR) in the stretches of river Bhagirathi, Farakka, West Bengal, India revealed that the landings of *N. Chitala* have declined rapidly (~70%) in recent decades. We are interested to understand the food chain dynamics of the species to find out the important parameters controlling the Chitala population in its natural habitat.

Based on the literature, we consider an intraguild predation (IGP) system consisting of Chitala (top predator), mugil (intermediate predator), and shrimp (basal prey). Two variants of IGP models governed by three coupled differential equations are considered for data modeling. In one model, Chitala is dependent only on mugil and shrimp. In the second model, there is an alternative food source available to chitala. The models are calibrated under the Bayesian modeling framework. Posterior estimates of the parameters for each model were obtained using the Gibbs algorithm. Reversible-jump Markov chain Monte Carlo method has been used to obtain the posterior model probabilities and the best model is selected using the posterior mode on the model space. The robustness of the statistical inference has been checked by considering different prior distributions.

Our finding suggests that the primary reason for unavailability is due to the high risk of extinction of mugil populations. Sensitivity analysis has confirmed that the biomass conversion rate from mugil to Chitala is the most important parameter. This study may be useful to develop management strategies for Chitala conservation by emphasizing on the regeneration of mugil populations.

Keywords: fisheries management; food chain; Intraguild predation; sensitivity analysis

P41

Adversarially Robust Learning:

A Generic Minimax Optimal Learner and Characterization

O. Montasser [1] , **S. Hanneke** [2] , **and N. Srebro** [1]

[1]*Toyota Technological Institute at Chicago*
[2] *Purdue University*

We present a minimax optimal learner for the problem of learning predictors robust to adversarial examples at test-time. Interestingly, we find that this requires new algorithmic ideas and approaches to adversarially robust learning. In particular, we show, in a strong negative sense, the suboptimality of the robust learner proposed by [1] and a broader family of learners we identify as local learners. Our results are enabled by adopting a global perspective, specifically, through a key technical contribution: the global one-inclusion graph, which may be of independent interest, that generalizes the classical one-inclusion graph due to [2]. Finally, as a byproduct, we identify a dimension characterizing qualitatively and quantitatively what classes of predictors H are robustly learnable. This resolves an open problem due to [1] and closes a (potentially) infinite gap between the established upper and lower bounds on the sample complexity of adversarially robust learning.

[1] O. Montasser, S. Hanneke, and N. Srebro. VC classes are adversarially robustly learnable, but only improperly, Proceedings of the Thirty-Second Conference on Learning Theory (2019).
[2] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting {0,1}-functions on randomly drawn points. Inf. Comput., (1994).

P42

The Hidden-Manifold Hopfield Model and a learning phase transition

Matteo Negri [1,2], **Clarissa Lauditi** [3,4], **Gabriele Perugini** [4], **Carlo Lucibello** [4] **and Enrico M. Malatesta** [4]

[1] *University of Rome 'La Sapienza', Department of Physics, Piazzale Aldo Moro 5, 00185 Roma, Italy*
[2] *CNR-NANOTEC, Institute of Nanotechnology, Rome Unit, Piazzale Aldo Moro, 00185 Roma, Italy*
[3] *Department of Applied Science and Technology, Politecnico di Torino, 10129 Torino, Italy*
[4] *Artificial Intelligence Lab, Bocconi University, 20136 Milano, Italy*

The Hopfield model has a long-standing tradition in statistical physics, being one of the few neural networks for which a theory is available. Extending the theory of Hopfield models for correlated data could help understand the success of deep neural networks, for instance describing how they extract features from data. Motivated by this, we propose and investigate a generalized Hopfield model that we name Hidden-Manifold Hopfield Model: we generate the couplings from $P=\alpha N$ examples with the Hebb rule using a non-linear transformation of $D=\alpha_D N$ random vectors that we call factors, with N the number of neurons. Using the replica method, we obtain a phase diagram for the model that shows a phase transition where the factors hidden in the examples become attractors of the dynamics; this phase exists above a critical value of α and below a critical value of α_D. We call this behaviour learning transition.

Preprint available at https://arxiv.org/abs/2303.16880

P43

Class-Imbalanced Classifiers: An Application in High Dimensional Datasets in Oncology

Gideon N. Nyakundi [12] **and John Ndiritu, Ph.D** [1]

[1] *Department of Mathematics University of Nairobi-Kenya*
[2] *Kajiado East Technical and Vocational College- Kenya*

Statistical analysis of imbalanced high dimensional data with missing data in oncology presents several challenges. One of the challenge is the classification problems for class imbalanced high-dimensional data. The standard classification methods that are often used for class-imbalanced data produce classification rules that do not accurately predict the minority class [1]. In the field of oncology, a comprehensive data analysis framework for datasets focusing on classification and prediction of survival of cancer patients is lacking. Random Forest Classifiers have comparatively outperformed other classifiers in handling disproportionate data distributions while providing great classification acccuracy. The proposed poster gives an extensive review of the application of Random Forest Classifiers in the prediction of minority class membership in class-imbalanced high dimensional data. The review incorporates the application of the techniques for modeling the performance classifier. This will demonstrate the practical utility of Random Forests Classifiers in Class-Imbalanced High Dimensional Data with a goal of encouraging their use in the statistical analysis and reporting cancer diagnosis and treatment routine studies.

Keywords: High Dimensional Data, Class Imbalance, Classification, Random Forests, Oversampling, Undersampling.

[1] Lin, W. J., & Chen, J. J. Briefings in bioinformatics. **14**, 1 (2013).

P44

Average case analysis of Lasso regression under ultra-sparse conditions

Koki Okajima, Xiangming Meng, Takashi Takahashi, Yoshiyuki Kabashima

Department of Physics, The University of Tokyo

An important concept in modern statistics is sparsity, where one assumes that only a few explanatory variables are sufficient in explaining the retrieved data. This allows the statistician to infer these from relatively few observations, even if the candidate variables are overwhelmingly abundant.

For instance, consider the *ultra-sparse* linear regression problem, where the observations $\boldsymbol{y} \in \mathbb{R}^M$ are given by a linear transformation of a large vector $\boldsymbol{x}_0 \in \mathbb{R}^N$ with only $d = O(1) \ll N$ non-zero entries:

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x}_0 + \boldsymbol{\xi}. \tag{1}$$

Here, $\mathbf{A} \in \mathbb{R}^{M\times N}$ is the sensing matrix, and $\boldsymbol{\xi} \in \mathbb{R}^M$ is the noise vector. Such situations are typical in applications such as material informatics, where the physical properties of a material is assumed to be explainable by a handful of features [1]. An important objective is then to recover without error the set of positions of non-zero components in x_0, which is often termed *support*, excluding as many unnecessary features in our interpretation as possible. A popular estimation method is the least absolute shrinkage and selection operator (Lasso) [2], which penalizes simple linear regression with the ℓ_1 norm of the vector:

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} \frac{1}{2M}\|\mathbf{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{x}\|_1. \tag{2}$$

Analytical techniques initially developed under disordered physics, such as the replica method, has been successful in predicting the Lasso solution's behavior when the matrix $\mathbf{A}$ and noise $\boldsymbol{\xi}$ is given randomly according to rather general distributions [3]. These analyses assume that the system dimensions (N, M, d) all diverge at the same rate while keeping their ratios constant. Our research provides a new way of applying the replica method for the ultra-sparse case, i.e. when $d = O(1)$, without assuming N and M to diverge at the same rate. This enhanced replica method allows us to predict the average properties of the Lasso solution, including its support recovery performance. Moreover, we provide a necessary condition for support recovery when the number of observations scale as $M = O(\log N)$. This complements the sufficient conditions given in previous literature [4].

References

[1] Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C., and Scheffler, M. Phys. Rev. E. **114**, 105503 (2015).

[2] Tibshirani, R. J. R. Stat. Soc., B: Stat. Methodol., **58**(1):267–288 (1990).

[3] Vehkaperä, M., Kabashima, Y., and Chatterjee, S. IEEE Trans. Inf. Theory. **62**(4):2100-2124.

[4] Dossal, C., Chabanol, M.-L., Peyré, G., and Fadili, J. Appl. Comput. Harmon. Anal. **33**(1):24-43.

P45

Evaluating the quality of pairwise maximum entropy models in large neural datasets

Valdemar K. Olsen[1] **and Yasser Roudi**[1]

[1]*Kavli Institute for Systems Neuroscience, Norwegian University of Science and Technology, Trondheim, Norway*

With rapid advancements in recording technology [1], much of computational neuroscience has effectively turned into describing large amounts of data as succinctly as possible [2]. One example of this is the abundance of different dimensionality reduction techniques applied in neuroscience [3]. Another, perhaps more interpretable, approach is to construct simple parametric models of the probability distribution over patterns of neuronal activity [4].

One model that has shown particular promise in capturing the experimentally observed probabilities of activity patterns, at least for few neurons N, is the pairwise maximum entropy model [e.g., 5–9]. However, whether this good performance for small N generalizes to larger N is unclear. Roudi *et al.* [10] performed a perturbative expansion in $N\bar{\nu}\delta t$, showing that the quality of the pairwise model should be linear in $N\bar{\nu}\delta t$ when $N\bar{\nu}\delta t \ll 1$ regardless of what the true distribution is, where $\bar{\nu}$ is the mean firing rate and δt is the binsize. Here, we investigate how the performance of the pairwise model scales with $N\bar{\nu}\delta t$ outside of this perturbative regime.

The performance of the pairwise model is typically measured as $G = 1 - d_{\text{pair}}/d_{\text{ind}}$, where d_{pair} and d_{ind} is the KL-divergence between the true and pairwise and true and independent distribution, respectively. This, however, is difficult to evaluate for large N because $d_{\text{pair}} = \sum_{\mathbf{s}}^{2^N} p_{\text{true}}(\mathbf{s}) \ln \frac{p_{\text{true}}(\mathbf{s})}{p_{\text{pair}}(\mathbf{s})}$ involves a sum over all 2^N states $\mathbf{s}$. We make two simplifications to nevertheless approximate G. First, we let p_{true} be the frequency distribution of the data p_{data}. Unsampled states would then have a probability of 0, which reduce the sum in d_{pair} to a more manageable size. Second, we approximate the partition function Z in the pairwise model as the $\hat{Z}$ that gets the unnormalized probabilities of the pairwise model as close to p_{data} as possible, only considering the sampled states. This approximation works by exploiting that a minority of the possible states carry the majority of the probability. $\hat{Z}$ allows us to successively evaluate p_{pair}, d_{pair}, and finally G.

We tested this approach on a Neuropixel dataset recorded from the visual, auditory, somatosensory, and motor cortices of freely moving rats [11]. In all cortical areas, we find that G is large for small $N\bar{\nu}\delta t$, consistent with previous findings, before it falls sharply and levels off at a non-zero G. This is the first systematic investigation of how well the pairwise model accounts for all correlations in neuronal data.

[1] I. H. Stevenson and K. P. Kording, Nature neuroscience **14**, 139 (2011).
[2] P. Gao and S. Ganguli, Current opinion in neurobiology **32**, 148 (2015).
[3] J. P. Cunningham and B. M. Yu, Nature neuroscience **17**, 1500 (2014).
[4] G. J. Stephens, L. C. Osborne, and W. Bialek, Proceedings of the National Academy of Sciences **108**, 15565 (2011).
[5] E. Schneidman, M. J. Berry, R. Segev, and W. Bialek, Nature **440**, 1007 (2006).
[6] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, and E. Chichilnisky, Journal of Neuroscience **26**, 8254 (2006).
[7] S. Yu, D. Huang, W. Singer, and D. Nikolić, Cerebral cortex **18**, 2891 (2008).
[8] E. Ganmor, R. Segev, and E. Schneidman, Journal of Neuroscience **31**, 3044 (2011).
[9] M. I. Chelaru, S. Eagleman, A. R. Andrei, R. Milton, N. Kharas, and V. Dragoi, Neuron **109**, 3954 (2021).
[10] Y. Roudi, S. Nirenberg, and P. E. Latham, PLoS computational biology **5**, e1000380 (2009).
[11] B. Mimica, T. Tombaz, C. Battistin, J. G. Fuglstad, B. A. Dunn, and J. R. Whitlock, bioRxiv (2022).

P46

Abstract template for Poster

S. Ariosto [12] , R. Pacelli [34], M. Pastore [5] , M. Gherardi [6] , F. Ginelli [12] and P. Rotondo [6]

[1] *Dipartimento di Scienza e Alta Tecnologia and Center for Nonlinear and Complex Systems, Università degli Studi dell'Insubria, Via Valleggio 11, 22100 Como, Italy*

[2] *I.N.F.N. Sezione di Milano, Via Celoria 16, 20133 Milano, Italy*

[3] *Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino, 10129 Torino, Italy*

[4] *Artificial Intelligence Lab, Bocconi University, 20136 Milano, Italy*

[5] *Université Paris-Saclay, CNRS, LPTMS, 91405 Orsay, France*

[6] *Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy*

Decades-long literature testifies to the success of statistical mechanics at clarifying fundamental aspects of deep learning. Yet the ultimate goal remains elusive: we lack a complete theoretical framework to predict practically relevant scores, such as the train and test accuracy, from knowledge of the training data. Huge simplifications arise in the infinite-width limit, where the number of units N in each hidden layer far exceeds the number P of training examples. This idealisation, however, blatantly departs from the reality of deep learning practice, where training sets are larger than the widths of the networks. Here, we show one way to overcome these limitations. The partition function for fully-connected architectures, which encodes information about the trained models, can be computed analytically with the toolset of statistical mechanics. The computation holds in the "thermodynamic limit" where both N and P are large and their ratio $\alpha = P/N$, which vanishes in the infinite-width limit, is now finite and generic. This advance allows us to obtain (i) a closed formula for the generalisation error associated to a regression task in a one-hidden layer network with finite α; (ii) an expression of the partition function (technically, via an "effective action") for fully-connected architectures with arbitrary number of hidden layers, in terms of a finite number of degrees of freedom (technically, "order parameters"); (iii) a demonstration that the Gaussian processes arising in the infinite-width limit should be replaced by Student-t processes; (iv) a simple analytical criterion to predict, for a given training set, whether finite-width networks (with ReLU activations) achieve better test accuracy than infinite-width ones. As exemplified by these results, our theory provides a starting point to tackle the problem of generalisation in realistic regimes of deep learning.

P47

New tools for sampling the posterior of neural networks

Giovanni Piccioli[1], **Emanuele Troiani**[1], **and Lenka Zdeborová**[1]

[1] *SPOC laboratory, EPFL*

Bayesian neural networks (BNN) offer an alternative learning paradigm to gradient based methods. On the other hand, Bayesian inference is usually impractical since it requires sampling from a complex, high dimensional posterior. In this work we study how to exactly sample from such a posterior using Markov chain Monte Carlo (MCMC) methods. We propose a new probabilistic model for BNN consisting of adding noise at every pre- and post- activation in the network, arguing that the resulting posterior is easier to sample using an efficient Gibbs sampler tailored to the probabilistic model we introduced. We show that such algorithm allows us to sample from the posterior of various architectures.

P48

Analysis of test error in the asymptotic polynomial regime for two-layer neural networks with respect to activation function

Anastasia Remizova[1], **Nicolas Macris**[1]

[1] *SMILS - IC - École Polytechnique Fédérale de Lausanne (EPFL), Switzerland*

Kernel models, such as the random features (RF) model [1] and the neural tangent kernel (NTK) model [2], may be regarded as approximations of two-layer neural networks. Particularly, they can be utilized for the high-dimensional asymptotic analysis of the generalization error, i.e., when the number of samples n, number of features d, and number of hidden-layer neurons N tend to infinity in suitable ways.

Recently, progress has been made in the asymptotic characterization of the performance of kernel ridge regression (KRR) in the polynomial regime [3] – that means, when the limit of n/d^K, $K \in \mathbb{N}$, is a positive constant δ_K. We use these results to investigate the neural network models of infinite width and derive precise conditions to observe the non-trivial behavior of the test error curve as a function of δ_K, such as the double descent phenomenon.

Moreover, with these findings, we examine how the activation function influences test error in the RF and NTK models. We analyze and compare possible qualitative pictures for several popular activations, such as ReLU and hyperbolic tangent, to draw conclusions on their theoretical performance.

[1] A. Rahimi, B. Recht, Random features for large-scale kernel machines. Advances in neural information processing systems, 1177–1184 (2008).

[2] A. Jacot, F. Gabriel, C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks. Advances in neural information processing systems, 8571–8580 (2018).

[3] H. Hu, Y. M. Lu, Sharp asymptotics of kernel ridge regression beyond the linear regime. arXiv:2205.06798 (2022).

P49

Abstract template for Poster in Youth in High-Dimensions

R.Rende[1], **F. Gerace**[1], **A. Laio**[1], **and S. Goldt**[2]

[1] *Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy*

Transformers are a type of neural network that achieves state-of-the-art performance in natural language processing, quantum physics, and structural biology. The key building block of transform- ers is a mechanism called self-attention, which is optimised to "fill-in-the-blank" in a given sequence, for example by predicting a missing word in a sentence. Despite their practical success, it remains unclear what and how self-attention learns from its data. Here, we give a precise analytical and numerical characterisation of self-attention trained on data drawn from a generalised Potts model with interactions between sites and Potts colours. We show that a single layer of self-attention with a small modification is able to to learn any two-body distribution exactly and efficiently through gradient descent, We compute the generalisation error of self-attention using the replica method from statistical physics and discuss connetions with pseudo-likelihood methods to solve the inverse Ising and Potts problem.

- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Advances in neural information processing systems 30 (2017).
- N. Bhattacharya, N. Thomas, R. Rao, J. Dauparas, P. K. Koo, D. Baker, Y. S. Song, and S. Ovchinnikov, bioRxiv:2020.12.21.423882 (2020).

P50

Microcanonical Hamiltonian Monte Carlo

Jakob Robnik[1] **and Uroš Seljak**[1,2]

[1]*Physics Department, University of California at Berkeley, Berkeley, CA 94720, USA*
[2]*Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

Microcanonical Hamiltonian Monte Carlo (MCHMC) [1, 2] is a class of models that follow fixed energy Hamiltonian dynamics, in contrast to Hamiltonian Monte Carlo (HMC) [4] which follows canonical distribution with different energy levels. MCHMC tunes the Hamiltonian function such that the marginal of the uniform distribution on the constant-energy-surface over the momentum variables gives the desired target distribution. It can also be viewed as an energy-preserving stochastic process [3], whose drift-diffusion discretization scheme is bias-free.

MCHMC exhibits favorable scaling with the condition number and its performance is independent of the dimensions. In high dimensional applications, this yields orders of magnitude speed-up relative to HMC, which scales as $\mathcal{O}(d^{-1/4})$. We have developed an efficient hyperparameter tuning scheme, which tunes the stepsize to achieve a desired energy error, ensuring a sufficiently low bias. By monitoring the autocorrelation, the momentum decoherence scale is tuned to be a fraction of the size of the typical set, which is in essence similar to the No-U-Turn condition.

We have tested MCHMC on the popular Bayesian inference problems and on the statistical physics of the scalar ϕ^4 field theory. It achieves an order of magnitude better performance than the state-of-the art methods like the HMC variant NUTS [5].

[1] J. Robnik, G. B. De Luca, E. Silverstein and U. Seljak, in review at JMLR (2022)
[2] G. Ver Steeg, A. Galstyan, Advances in Neural Information Processing Systems **34** (2021).
[3] J. Robnik and U. Seljak, Microcanonical Langevin Monte Carlo, in review at PRL (2023)
[4] S. Duane, A. Kennedy, B. Pendleton and D. Roweth, Physics letters B **21**, 195 (1987).
[5] M. Hoffman, A. Gelman, et al., J. Mach. Learn. Res. **1**, 15 (2014).

P51

A Unifying Approach for Statistical Inference in High-Dimensional Generalized Linear Models

Kazuma Sawaya[1], **Yoshimasa Uematsu**[2], **and Masaaki Imaizumi**[1]

[1]*The University of Tokyo*
[2]*Hitotsubashi University*

The generalized linear model (GLM) is one of the most popular statistical models to capture the nonlinear relationships between features and a target variable. In recent years, high-dimensional GLMs are important in many applications for the inference of high-dimensional data. However, in a high-dimensional limit where both the sample size and the dimension are large and their ratio is fixed, several theoretical studies [1,2] have shown that maximum likelihood estimators of high-dimensional GLMs have asymptotic biases. This is a serious problem because it can give rise to erroneous scientific findings. This presentation aims to provide a unified framework for estimating such asymptotic biases to correct classical statistical inferences.

Suppose that we observe n samples of a pair of a feature vector $X_i \in \mathbb{R}^p$ and a target variable $Y_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$. We denote a design matrix $X = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times n}$ and a target vector $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$. A generalized linear model assumes that

$$\mathbb{E}[Y_i | X_i] = g'\left(X_i^\top \beta^\circ\right), \tag{1}$$

where we call given g' an inverse link function which is the derivative of $g : \mathbb{R} \to \mathbb{R}$ and captures the nonlinearity, and $\beta^\circ \in \mathbb{R}^p$ is the unknown coefficient vector whose coordinates measure the marginal dependency of each feature.

We will focus on the inference on each coefficient $\beta_j^\circ, j = 1, \ldots, p$ in moderately high dimensions, where the number of observations n grows proportionally with the number of features p with $p/n \to \kappa \in (0, \infty)$.

[1] Sur, P. and Candès, E.J., 2019. A modern maximum-likelihood theory for high-dimensional logistic regression. Proceedings of the National Academy of Sciences, 116(29), pp.14516-14525.

[2] Feng, O.Y., Venkataramanan, R., Rush, C. and Samworth, R.J., 2022. A unifying tutorial on approximate message passing. Foundations and Trends ® in Machine Learning, 15(4), pp.335-536.

P52

AutoencODEs: a control theoretic framework for modeling Autoencoders

A. Scagliotti[1,2], **C. Cipriani**[1], **and M. Fornasier**[1,2]

[1]*Technical University of Munich, Germany*
[2]*Munich Center for Machine Learning (MCML), Germany*

In 2017, it was observed that Residual Neural Networks (ResNets) can be studied as discretizations of continuous-time control systems $\dot{x}(t) = F(x(t), u(t))$, where $u = u(t)$ is the control function that drives the dynamics. This kind of systems are often called in literature NeurODEs, and, in the last years, Control Theory has been fruitfully applied to study the properties of existing ResNets, and to develop new ones.
Since the dimension of the phase-space of a NeurODE is constant, they have not been used so far to model Deep Learning architectures where the dimensions of the inputs and the outputs vary along the layers. In particular, this is the case for Autoencoders, where the dimension of the data is compressed during the encoding phase, and it increases during the decoding.
In this poster, we describe a NeurODE of the form $\dot{x}(t) = F(t, x(t), u(t))$, aimed at modeling a continuous-time Autoencoder. Here, the explicit dependence of the dynamics on the time variable t is crucial during the encoding phase, when we reduce progressively the motion to lower-dimensional subspaces. With a similar strategy, in the decoding part we manage to gradually restore the original dimension of the data.
Finally, we employ a Mean-Field argument to study the stability of the training process of our NeurODE when the size of the data set tends to infinity.
Based on an ongoing project with C. Cipriani and M. Fornasier.

P53

From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks

Luca Arnaboldi[1], **Ludovic Stephan**[1], **Florent Krzakala**[1] **and Bruno Loureiro**[2]

[1]*(Presenting author underlined) École Polytechnique Fédérale de Lausanne (EPFL), IdePHICS Lab, CH-1015 Lausanne, Switzerland*

[2]*Département d'Informatique, École Normale Supérieure (ENS) - PSL & CNRS, F-75230 Paris cedex 05, France*

We investigate the one-pass stochastic gradient descent (SGD) dynamics of a two-layer neural network trained on Gaussian data and labels generated by a similar, though not necessarily identical, target function. We rigorously analyse the limiting dynamics via a deterministic and low-dimensional description in terms of the sufficient statistics for the population risk. Our unifying analysis bridges different regimes of interest, such as the classical gradient-flow regime of vanishing learning rate, the high-dimensional regime of large input dimension, and the over-parameterised "mean-field" regime of large network width, covering as well the intermediate regimes where the limiting dynamics is determined by the interplay between these behaviours. In particular, in the high-dimensional limit, the infinite-width dynamics is found to remain close to a low-dimensional subspace spanned by the target principal directions. Our results therefore provide a unifying picture of the limiting SGD dynamics with synthetic data.

[1] Arnaboldi, L., Stephan, L., Krzakala, F. & Loureiro, B. From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. (arXiv,2023), https://arxiv.org/abs/2302.05882

P54

Data-driven separation between feature and lazy learners for higher-order statistics

Eszter Székely[1], **Federica Gerace**[1], **Lorenzo Bardone**[1] **and Sebastian Goldt**[1]

[1]*Scuola Internazionale Superiore di Studi Avanzati, SISSA*

We compare shallow two-layer networks in the fully-trained and lazy learning regime on tasks that rely on the information contained in the higher-order statistics of the inputs. We design synthetic models of data where we control the relative importance of higher-order cumulants and study in which settings end-to-end trained networks achieve better performance than random features. We further study the features of data that neural networks fit.

P55

The Amount of Data Needed to Train Dense Hopfield Networks

Robin Thériault[1] **and Daniele Tantari**[2]

[1]*Scuola Normale Superiore, Pisa*
[2]*University of Bologna*

Hopfield networks with 2-body interactions [1] became popular because they could serve as simple analytically tractable models for both biological and artificial neural networks. However, it was soon realized that 2-body Hopfield networks could only store a number of patterns M proportional to their dimensionality N [2]. On the other hand, it was also discovered that p-body interactions allowed them to store up to $M \propto N^{p-1}$ patterns [3]. This idea was recently rediscovered thanks to a feedforward implementation of p-body Hopfield networks [4], the so-called dense Hopfield netwoks, which were shown to be resistant to adversarial attacks [5]. With their rise in popularity, it becomes important to study other characteristics of p-body Hopfield networks, notably the amount of data needed to train them. We explore the latter subject analytically by using the replica method in the teacher-student setting. This approach allows us to compute the phase diagram as a function of the number of data points and the amount of noise they contain. In particular, we show that beyond a certain level of noise, the student network needs at least $M \propto N^{p-1}$ training examples to learn a pattern held by the teacher. We also investigate the $p \to \infty$ limit of the phase diagram. Our results are compared and contrasted with Monte-Carlo simulations.

References

[1] J. J. Hopfield, *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Physical Review Letters*, vol. 55, no. 14, p. 1530, 1985.

[3] E. Gardner, *Journal of Physics A: Mathematical and General*, vol. 20, no. 11, p. 3453, 1987.

[4] D. Krotov and J. J. Hopfield, *Advances in neural information processing systems*, vol. 29, 2016.

[5] D. Krotov and J. Hopfield, *Neural computation*, vol. 30, no. 12, pp. 3151–3167, 2018.

P56

On the largest canonical correlation coefficients between the past and the future of a high-dimensional time series

D. Tieplova[1], **P. Loubaton**[2], **and J.Yao**[3]

[1]*The Abdus Salam International Centre for Theoretical Physics, Italy*
[2]*Université Gustave Eiffel Laboratoire d'Informatique Gaspard Monge, France*
[3]*The Chinese University of Hong Kong (Shenzhen) School of Data Science, China*

Large random matrix theory became a powerful tool in the domain of high dimensional statistical signal processing, where traditionally we observe double asymptotic regime in which the dimension of the time series and the sample size both grow towards infinity. In this work we address less studied structure of large random matrices, and use the corresponding results to address an important statistical problem. More specifically, we consider a M – dimensional multivariate time series $(y_n)_{n\in\mathbb{Z}}$ generated as $y_n = v_n + u_n$, where v_n is "noise", centred complex Gaussian vector such that $\mathbb{E}(v_n v_{n+k}^*) = R\delta_k$ and u_n is non observable "information" part which admits causal state space representations

$$\begin{aligned} x_{n+1} &= Ax_n + B\omega_n \\ u_n &= Cx_n + D\omega_n \end{aligned}$$

here $(\omega_n)_{n\in\mathbb{Z}}$ – $K \leq M$ dimensional white noise sequence $\mathbb{E}(\omega_{n+k}\omega_n^*) = I_K\,\delta_k$; $(x_n)_{n\in\mathbb{Z}}$ is a P-dimensional Markovian sequence; A is a deterministic $P \times P$ matrix whose spectral radius is strictly less than 1; B, C, D are deterministic matrices of sizes $P \times K$, $M \times P$ and $M \times K$ respectively.

An important statistical problem is the estimation of the minimal dimension P of the state space representations of u from N samples $y_1, \ldots, y_N$. If L is any integer larger than P, the traditional approaches are based on the observation that P coincides with the number of non zero singular values of matrix C^L whose entries are inner products between the elements of orthonormal bases of the "future"(generated by the components of $y_{n+L}, \ldots, y_{n+2L-1}$) and the "past" (generated by the components of $y_n, \ldots, y_{n+L-1}$). In the low-dimensional regime where $N \to +\infty$ while M and L are fixed, the matrix C^L can be consistently estimated by its empirical $\hat{C}^L = (Y_f Y_f^*)^{-1/2} Y_f Y_p^* (Y_p Y_p^*)^{-1/2}$ where

$$Y_p^{(L)} = \begin{pmatrix} y_1 & y_2 & \cdots & y_N \\ y_2 & y_3 & \cdots & y_{N+1} \\ \vdots & \vdots & \vdots & \vdots \\ y_L & y_{L+1} & \cdots & y_{N+L-1} \end{pmatrix}, \; Y_f^{(L)} = \begin{pmatrix} y_{L+1} & y_{L+2} & \cdots & y_{N+L} \\ y_{L+2} & y_{L+3} & \cdots & y_{N+L+1} \\ \vdots & \vdots & \vdots & \vdots \\ y_{2L} & y_{2L+1} & \cdots & y_{N+2L-1} \end{pmatrix},$$

and P can be evaluated from the largest singular values of $\hat{C}^L$. If however $\frac{ML}{N} = c_N \to c_* \in (0, 1]$, L being fixed, the above estimate $\hat{C}^L$ do not converge towards its value in the spectral norm sense. It is therefore not obvious whether the largest singular values of $\hat{C}^L$ can be used in order to estimate P consistently. In this work we study the behaviour of the singular values of $\hat{C}^L$ in the above high-dimensional regime. It is shown that in the case where $u = 0$, or equivalently $y = v$, the empirical singular values distribution of $\hat{C}^L$ converge towards a limit with explicit form with support $[0, 2\sqrt{c_*(1-c_*)}] \cup \{1\}\mathbf{1}_{c_* > \frac{1}{2}}$. When u is present, as expected, some singular values of $\hat{C}^L$ escapes the limiting support. In a simpler case we can directly relate the number of outliers to parameters such as eigenvalues of A and variance of v_n. Also we show that in general case P coincides with the number of singular values of $\hat{C}^L$ that are larger than $2\sqrt{c_*(1-c_*)}$, provided $c_* < \frac{1}{2}$, and the signal u is powerful enough compared to the noise and the non zero singular values of C^L are large enough.

P57

How deep convolutional neural networks lose spatial information with training

Umberto M. Tomasini[1,*], **Leonardo Petrini**[1,*], **Francesco Cagnetta**[1] **and Matthieu Wyart**[1]

[1] *Institute of Physics, École Polytechnique Fédérale de Lausanne (EPFL).*

A central question of machine learning is how deep nets manage to learn tasks in high dimensions. An appealing hypothesis is that they achieve this feat by building a representation of the data where information irrelevant to the task is lost. For image datasets, this view is supported by the observation that after (and not before) training, the neural representation becomes less and less sensitive to diffeomorphisms acting on images as the signal propagates through the net. This loss of sensitivity correlates with performance, and surprisingly correlates with a *gain* of sensitivity to white noise acquired during training. These facts are unexplained, and as we demonstrate still hold when white noise is added to the images of the training set. Here, we *(i)* show empirically for various architectures that stability to image diffeomorphisms is achieved by both spatial and channel pooling, *(ii)* introduce a model scale-detection task which reproduces our empirical observations on spatial pooling and *(iii)* compute analitically how the sensitivity to diffeomorphisms and noise scales with depth due to spatial pooling. The scalings are found to depend on the presence of strides in the net architecture. We find that the increased sensitivity to noise is due to the perturbing noise piling up during pooling, after being rectified by ReLU units.

[1] Tomasini, Petrini, Cagnetta, Wyart, ”How deep convolutional neural networks lose spatial information with training”, arXiv:2210.01506, accepted in the Workshop ”Physics4AI”, ICLR 2023.

P58

Emergent representations in networks trained with the Forward Forward algorithm

Niccolò Tosato[1,2,*], **Lorenzo Basile**[2,3,*], **Emanuele Ballarin**[2]

Giuseppe de Alteriis[4], **Alberto Cazzaniga**[1], **and Alessio Ansuini**[1]

[1] *AREA Science Park, Trieste (Italy)*
[2] *University of Trieste, Trieste (Italy)*
[3] *The Abdus Salam International Centre for Theoretical Physics, Trieste (Italy)*
[4] *King's College London, London (UK)*

Deep learning is a highly effective way to create artificial intelligence, with tremendous implications for science, technology and culture. At the core of deep learning is the backpropagation (BP) algorithm [4], that efficiently compute the gradients necessary to optimize the parameters of the network. However, BP lacks biological plausibility, leading to many attempts to address this issue. The most recent of these efforts, the Forward Forward (FF) algorithm developed by G.Hinton [1], eliminates the need to propagate error derivatives and store neural activities. In a standard classification context, the application of FF requires the designation of *positive* and *negative* data. For instance, to classify images, one could assign positive (negative) data to those images which have the correct (incorrect) classification embedded via one-hot encoding at the border. The FF algorithm then attempts to differentiate between positive and negative data by optimizing a goodness function (such as, e.g., the L_2 norm of the activations), following an approach similar to contrastive learning. Results have been observed to be satisfactory for the MNIST classification task [1], a well-known supervised learning benchmark.
This work goes beyond performance, investigating the strong similarities between biological and artificial neuron ensembles. Our experiments demonstrate sparseness in representations learned using FF, and similarities to cortical ensembles found in early stages of sensory processing [2].
Neurons that form ensembles are highly specialized and are found to activate selectively with a high probability when presented with positive data, while only exhibiting minimal activation to negative data or unexpected stimuli. Preliminary results indicate that there may be collective suppression mechanisms similar to those of inhibitory neurons [5].
Although optimizing the standard Cross Entropy loss does not appear to produce the behavior we observe, this may not be solely due to the FF algorithm. In fact, similar results can be obtained by replacing FF with backpropagation while using the same training procedure. This suggests that the focus should be on the purpose and biological meaning of the loss function rather than the training algorithm [3].

[1] Geoffrey Hinton, "The forward-forward algorithm: Some preliminary investigations", arXiv preprint arXiv:2212.13345 (2022).
[2] Jae-eun Kang Miller, Inbal Ayzenshtat, Luis Carrillo-Reid, and Rafael Yuste, "Visual stimuli recruit intrinsically generated cortical ensembles", Proceedings of the National Academy of Sciences, 111(38):E4053–E4061 (2014).
[3] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Chris- tensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al., "A deep learning framework for neuroscience", Nature neuroscience, 22(11):1761–1770 (2019).
[4] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams, "Learning representations by back-propagating errors", Nature, 323(6088):533–536 (1986).
[5] Rafael Yuste, "From the neuron doctrine to neural networks", Nature reviews neuroscience, 16(8):487–497 (2015).

P59

On the Leveraging of MLab Facilities to Mine the Nigerian Internet Traffic Measurement

Christopher Godwin Udomboso* and Margaret Esther Oluwadare

Computational Statistics Unit, Department of Statistics, University of Ibadan, Ibadan
cg.udomboso@gmail.com
margaretoluwadare@outlook.com

* Corresponding Author

Abstract: Measurement Lab (M-Lab) is an open source project that provides data on Internet performance measurements facilities. It came into existence in 2009 from a crucial discussion by Vint Cerf at the New America's Open Technology Institutes. This study leverages on the facilities provided by M-Lab to explore some major providers in the Nigerian telecommunication services providers. The data obtained were Internet Network Metrics (INM) observed in Globalcom, Airtel and Smile networks in Ibadan metropolis over a 9-month period. An R code was written to interact and mined the observed INM. Results indicated that the preferred data service provider is Airtel Nigeria, which had larger network readings, better latency, upload and download metrics. It was also observe that the Nigerian data network performs at its lowest at peak periods. Further studies are needed on the nature of data collected, perform deep learning and suggest possible decisions that would enhance the data service of Nigeria.

Keywords: Internet traffic, open source, internet network metrics, M-Lab

P60

The geometry of hidden representations of large transformer models

Lucrezia Valeriani[1,2] **Diego Doimo**[3] **Francesca Cuturello**[1] **Alessandro Laio**[3,4]
Alessio Ansuini[1,*] **Alberto Cazzaniga**[1,*]

[1] AREA Science Park, Padriciano 99, 34149 Trieste, Italy
[2] University of Trieste, Trieste 34127, Italy
[3] SISSA, via Bonomea 265, 34136 Trieste, Italy
[4] ICTP, Strada Costiera 11, 34014 Trieste, Italy

[*] **Correspondence:**
alessio.ansuini@areasciencepark.it
alberto.cazzaniga@areasciencepark.it

Large transformers are powerful architectures for self-supervised analysis of data of various nature, ranging from protein sequences to text to images. In these models, the data representation in the hidden layers live in the same space, and the semantic structure of the dataset emerges by a sequence of functionally identical transformations between one representation and the next. We here characterize the geometric and statistical properties of these representations, focusing on the "evolution" of such proprieties across the layers. By analyzing geometric properties such as the intrinsic dimension (ID) and the neighbor composition we find that the representations evolve in a strikingly similar manner in transformers trained on protein language tasks and image reconstruction tasks. In the first layers, the data manifold "expands", becoming high-dimensional, and then it contracts significantly in the intermediate layers. In the last part of the model, the ID remains approximately constant or forms a second shallow peak. We show that the semantic complexity of the dataset emerges at the end of the first peak. This phenomenon can be observed across many models trained on diverse datasets. Based on these observations, we suggest using the ID profile as an unsupervised proxy to identify the layers which are more suitable for downstream learning tasks.

P61

An Application of Correlation-Based Clustering for data imputation

Diego Velásquez Varela [1]

[1] (dvelasquev@eafit.edu.co)

Missing data is a common challenge in various fields, impacting the accuracy and reliability of analyses. Data imputation methods aim to address this issue by estimating missing values based on available information. This project presents a novel approach to data imputation, leveraging the Correlation Connected Clusters (4C) method, which combines Principal Component Analysis (PCA) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The 4C method identifies patterns and correlations within datasets, making it a promising foundation for a new data imputation methodology.

The proposed approach utilizes PCA to reduce dimensionality, allowing for efficient data analysis and noise reduction. By transforming the data into a lower-dimensional space, the method preserves the most relevant information while discarding noise. Next, DBSCAN is employed to identify clusters and outliers, providing a robust framework for correlation-based clustering.

The main advantage of the proposed method is its robustness against noise, which is crucial for accurate data imputation. By combining the strengths of PCA and DBSCAN within the 4C framework, this approach identifies and preserves meaningful patterns and correlations while mitigating the impact of noise on imputation results.

The long-term objective of this project is to develop a comprehensive data imputation methodology based on the 4C method. With its unique combination of PCA and DBSCAN, this approach has the potential to outperform traditional imputation methods and significantly advance the field of high-dimensional data analysis.

[1] C. Böhm, K. Kailing, P. Kröger, A. Zimek, Computing Clusters of Correlation Connected Objects, Proc. 2004 ACM SIGMOD Int. Conf. Management of Data, 455-466 (2004).
[2] M. M. Breunig, H. P. Kriegel, R. T. Ng, J. Sander, LOF: Identifying Density-Based Local Outliers, SIGMOD Rec. 29, 93 (2000).
[3] E. P. Kastampolidis, P. G. Sarigiannidis, I. D. Moscholios, G. I. Papadimitriou, The 4C Method: Clustering Based on Correlations and Similarities, IEEE Trans. Knowl. Data Eng. 31, 1900 (2019).
[4] J. C. Bezdek, R. J. Hathaway, VAT: A Tool for Visual Assessment of (Cluster) Tendency, Proc. IEEE Intl. Joint Conf. Neural Netw. 3, 2225 (2002).
[5] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: An Efficient Data Clustering Method for Very Large Databases, SIGMOD Rec. 25, 103 (1996).
[6] S. Ramaswamy, R. Rastogi, K. Shim, Efficient Algorithms for Mining Outliers from Large Data Sets, SIGMOD Rec. 29, 427 (2000).

P62 **Enrico Ventura**, poster abstract:

The role of noise in helping to learn is nowadays a well consolidated concept in the field of artificial neural networks that is likely to open new perspectives in the understanding of several brain functions, especially for what concerns memory formation and consolidation. In this matter, the training-with-noise algorithm proposed by Gardner and collaborators stood out as a valid alternative to linear perceptrons by employing a set of noisy independently generated configurations to train recurrent networks that retrieve and generalize random data. Nevertheless the success of this procedure appears limited to low levels of noise and moderate memory loads. We show how inserting an internal structure in noisy training data can substantially improve the memory performance of the network. It is also proved, by means of simple analytical arguments supported by numerics, that the widely celebrated unlearning rule coincides with the training-with-noise algorithm when noise is maximal and data are fixed points of the dynamics. Furthermore, the optimal structure of noise is analyzed in relation with particular positions in the landscape of attractors and used to build a sampling scheme for noisy configurations that allows to outperform both the training-with-noise and the unlearning procedures.

P63

Abstract template for Poster in Youth in High-Dimensions

L.L.Viteritti[2], **R.Rende**[1], **F.becca**[2]

[1] *Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, Italy*

[2] Dipartimento di Fisica, Universit'a di Trieste, Strada Costiera 11, I-34151 Trieste, Italy

The Transformer architecture has become the state-of-art model for natural language processing tasks and, more recently, also for computer vision tasks, thus defining the Vision Transformer (ViT) architecture. The key feature is the ability to describe long-range correlations among the elements of the input sequences, through the so-called self-attention mechanism. Here, we propose an adaptation of the ViT architecture with complex parameters to define a new class of variational neural-network states for quantum many-body systems, the ViT wave function. We apply this idea to the onedimensional $J_1 - J_2$ Heisenberg model, demonstrating that a relatively simple parametrization gets excellent results for both gapped and gapless phases. In this case, excellent accuracies are obtained by a relatively shallow architecture, with a single layer of self-attention, thus largely simplifying the original architecture. Still, the optimization of a deeper structure is possible and can be used for more challenging models, most notably highly-frustrated systems in two dimensions. The success of the ViT wave function relies on mixing both local and global operations, thus enabling the study of large systems with high accuracy.

- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Advances in neural information processing systems 30 (2017).
- G. Carleo and M. Troyer, Science 355, 602 (2017).

P64

Study on thermodynamics and kinetics of nucleic acid base pairs by statistical physics

Yujie Wang [1] , and Taigang Liu [2]

[1]Department of Physics, Zhoukou Normal University, Zhoukou 466000, China
[2] SChool of Medical Engineering, Xinxiang Medical University, Xinxiang 453003, China

According to statistical physics, firstly, we study the thermodynamics and kinetics of single base pair of nucleic acid molecules RNA and DNA. In thermodynamics, according to the simulated trajectory at different temperatures, we attained the enthalpy and entropy changes between the open and closed states of single base pair, which are consistent with the parameters of the nearest neighbour model in experiment; In kinetics, the average lifetime of the closed state and open state were obtained, and they showed different temperature dependences[1]. Secondly, The thermodynamic parameters and the opening-closing switching characteristic of single base pair were obtained at different ions concentration. In thermodynamics, the enthalpy change and the entropy change of the base pair were obtained at different ions concentration. It was found that the enthalpy is not effected by the ions concentration, however, the entropy is effected and decreases with decreased ions concentration, which is consistent with experimental results. In kinetics, according to the average lifetime of open state and closed state, it was found that the average lifetime of closed state were not vary with the different ions concentration, however, the average lifetime of the open state increases with the decrease of ion concentration, which is consistent with the experimental results[2]. Finally, by studying the thermodynamics and kinetics of the opening and closing of the terminal base pair of different nearest neighbour base, we can see that the nearest neighbour base not only affects the thermodynamics parameters, that is, the enthalpy and entropy changes of the base pair during switching, but also affects the average lifetime of the base pair in each conformational state and the conversion rate between them in kinetics[3].

[1] Y. Wang, T. Liu, T. Yu, Z. Tan, and W. Zhang, RNA. **26**, 470 (2020).
[2] Y. Wang, Z. Wang, Y. Wang, T. Liu, and W. Zhang, J. Chem. Phys. **148**, 045101 (2018).
[3] Y. Wang, S. Gong, Z. Wang, and W. Zhang, J. Chem. Phys. **144**, 115101, (2016).

P65

Attributed Graph Alignment: Fundamental Limits and Efficient Algorithms

Lele Wang[1]

[1]*University of British Columbia*

We consider the graph alignment problem, where the goal is to identify the vertex/user correspondence between two correlated graphs. Existing work mostly recovers the correspondence by exploiting the user-user connections. However, in many real-world applications, additional information about the users, such as user profiles, might be publicly available. In this talk, we introduce the attributed graph alignment problem, where additional user information, referred to as attributes, is incorporated to assist graph alignment. We establish both the information-theoretic limits and the feasible region by polynomial-time algorithms for the attributed graph alignment. Our results span the full spectrum between models that only consider user-user connections and models where only attribute information is available.

[1] Ziao Wang, Ning Zhang, Weina Wang, and Lele Wang, "On the feasible region of efficient algorithms for attributed graph alignment," *Proceedings of the IEEE International Symposiums on Information Theory (ISIT),* Espoo, Finland, July 2022.

[2] Ning Zhang, Weina Wang, and Lele Wang, "Attributed graph alignment," *Proceedings of the IEEE international symposium on information theory (ISIT),* Melbourne, Australia, July 2021.