UNLOCKING THE POTENTIAL OF DATA WITH MATHEMATICS

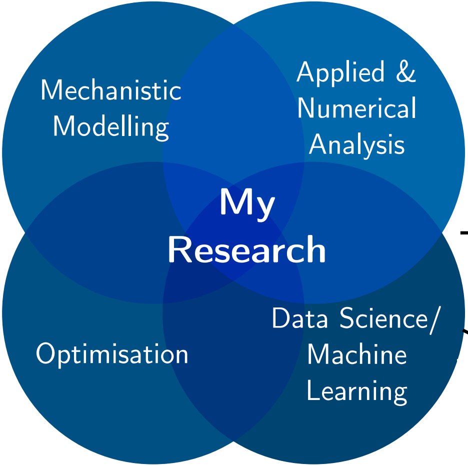
Lisa Maria Kreusser

University of Bath

VC Research Day, University of Bath

February 6, 2024

Unified framework for unlocking the full potential of data

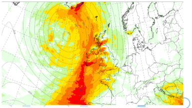


Biology → fingerprints, networks

Engineering →

textile production

Climate Science → power systems, weather prediction



Goal: Structures across scales

MICROSCOPIC FORMULATION
agent-based models (high-dimensional equations)
for properties (e.g. positions) of large number of agents

↓ rigorous bridge

MACROSCOPIC FORMULATION
mean-field equations (low-dimensional equations)
for macroscopic structures (e.g. density functionals)

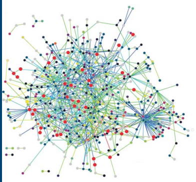
separate approaches

Originality:

- **New perspective:** bridge micro- and macroscopic description
- **Importance:**
  - probe microscopic system via macroscopic observables
  - mathematical formulations using differential equations and optimisation of energies allow the development of different methods in applications
  - rigorous bridge implying reliability of approaches

Examples of structures across scales in biology/medicine

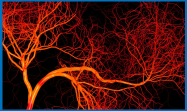
INTERACTION NETWORKS



e.g. molecular networks, social networks

APPLI-CATIONS: Biology, Medicine, Society

TRANSPORT NETWORKS



e.g. leaf venation, blood circulation

(P1) Construction of weighted graph

Given $N + M$ data points $V = \{X_1, \dots, X_{N+M}\}$:

- Determine **similarity measure** $w_{i,j}$ between data points X_i and X_j
- **Graph construction** based on similarity measure
- **Partition of the graph** using mathematical models[1]

(a) Data set

(b) Graph $G = (V, w)$

[1] Dunbar, Elliott, LMK, 2022

(P2) Mumford-Shah model for image segmentation[2]

- **Diverse applications** of image segmentation

- **Mumford-Shah model:** Minimisation of the energy functional

$$\mathcal{E}^{MS}(C,u) = \int_{\Omega} (u-u_0)^2 \,\mathrm{d}x + \mu \int_{\Omega\setminus C} |\nabla u|^2 \,\mathrm{d}x + \nu |C|$$

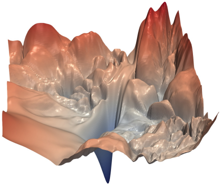
for fixed parameters $\mu, \nu > 0$

- **Ambrosio-Tortorelli approximation** of $|C|$
  - one of the most computationally efficient approximations
  - uses the Ginzburg-Landau functional $\mathbb{E}_{\epsilon}^{GL}(v) = \int_{\Omega} \epsilon |\nabla v|^2 + \frac{1}{\epsilon} W(v) \,\mathrm{d}x$ for double well potential $W\colon \mathbb{R} \to [0,+\infty)$ and $\epsilon > 0$
- **Convergence of minimisers** for important approximative model combining techniques from a variety of fields of mathematics

[2] I. Fonseca, LMK, C.-B. Schönlieb and M. Thorpe, IUMJ, 2023.

Many machine learning models are non-convex:

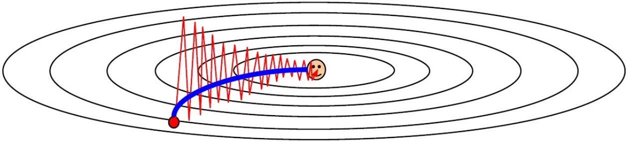
- K-means clustering, Deep Learning, ...



Credit: Li et al., arXiv:1712.09913

Aim: Developed modified Laplacian Smoothing Gradient Descent[3]

- Circumvent sharp minima and saddle points
- Avoid slow progress in shallow directions



Credit: Osher et al.

[3] LMK, Osher, Wang; EJAM, 2023

(P4) Score-based diffusion models

- Surpassed GANs in many generation tasks
- Theoretical foundations:
  - stochastic differential equations (SDEs)
  - probability flows, given by ordinary differential equations (ODEs)
  - mean-field dynamics for densities, given by Fokker-Planck equations

⇒ **Investigation of SDE-ODE gap via Fokker-Planck equation** [4]

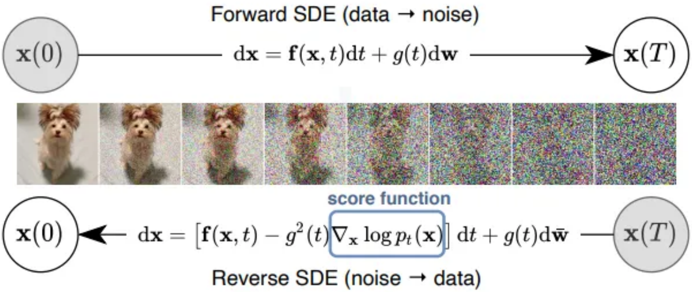


Image source: Song et al., https://arxiv.org/pdf/2011.13456.pdf

[4] T. Deveney, J. Stanczuk, LMK, C. Budd and C.-B. Schönlieb, arXiv:2311.15996

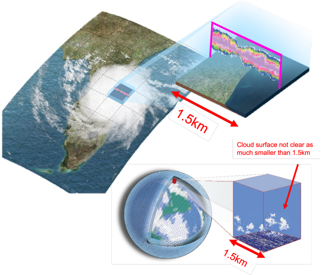
ICMS Knowledge Exchange Fellowship (09/2023 - 08/2024)

Example: Data-driven modelling of cloud organisation

- Best existing models have a resolution of around 1.5 to 20km
- Cloud fraction and cloud perimeter can be computed using physical laws, albeit are computationally expensive

⇒ Development of data-driven approaches using temperature, humidity and turbulence data

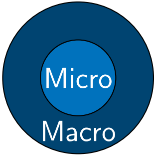
In collaboration with Prof. C. Budd (Bath), Dr K. Van Weverburg (Royal Met. Institute of Belgium) and Dr C. Morcette (Met Office)



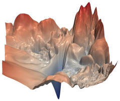
Images courtesy of Kwinten Van Weverburg

Conclusion: Four overarching topics

1. **Applied Analysis and differential equations** for establishing theoretical guarantees



2. **Numerical analysis and optimisation** including the development of new numerical methods and optimisation algorithms

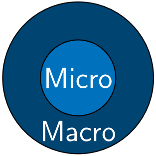


3. **Mathematical foundations of data science/machine learning** including semi-supervised learning and generative modelling

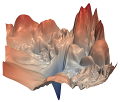
4. **Translation to applications** in biology, engineering and climate science

$\Rightarrow$ **Breadth of research in the mathematics of machine learning building the foundations for substantial progress in applications**

Conclusion: Four overarching topics



1. **Applied Analysis and differential equations** for establishing theoretical guarantees



2. **Numerical analysis and optimisation** including the development of new numerical methods and optimisation algorithms

3. **Mathematical foundations of data science/machine learning** including semi-supervised learning and generative modelling

4. **Translation to applications** in biology, engineering and climate science

⇒ **Breadth of research in the mathematics of machine learning building the foundations for substantial progress in applications**

THANK YOU VERY MUCH FOR YOUR ATTENTION!