# Mathematical Methods in Data Science (M24)

**George Stepaniants, Tram Nguyen, Alexei Shadrin**

Data-driven methods are the dominant tools by which we collect and analyse information in the real world. This course will introduce the three mathematical pillars that allow us to study the convergence, stability, and generalisation of data-driven estimators. Section I focuses on *statistical learning theory*, which studies how estimators learned from empirical data converge as the number of samples increases, and how their predictions generalise beyond the observed data. Section II builds the foundations of *inverse problems*, asking when accurate prediction or small risk implies accurate estimation of underlying parameters, functions, or signals, and what structural assumptions or regularisation strategies are needed for this to hold. Lastly, Section III gives an overview of *approximation theory*, which studies good and best approximations of important function classes, such as continuous, integrable, and smooth functions, and quantifies the convergence of approximation schemes. We list possible subtopics that each section of the course will cover:

**Section I: Statistical Learning Theory** (Lecturer: Stepaniants)

- Introduction to empirical risk minimisation
- Complexity, concentration, and generalisation
- Rates of estimation and minimax theory

**Section II: Theory of Inverse Problems** (Lecturer: Nguyen)

- Inverse problems: ill-posedness and regularisation
- Variational formulations, regularisers, and solution methods
- Bayesian estimation, inversion, priors
- Data-driven methods and modern applications

**Section III: Approximation Theory** (Lecturer: Shadrin)

- Good and best approximation in normed function spaces
- Polynomial approximation, positive linear operators, trigonometric approximation
- Introduction to wavelets, multiresolution analysis, Daubechies orthogonal wavelets

### Prerequisites

All mathematical theory will be introduced in a self-contained way but students are expected to have a working understanding of mathematical and functional analysis and probability theory. Prior exposure to scientific computing and optimisation is advantageous. The Part II courses *Linear Analysis, Statistical Modelling, Principles of Statistics, and Probability and Measure* are most relevant, and any crucial topics arising in these courses will be reviewed.

**Literature**

- General textbooks:
  1. M.J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint,* Vol. 48. Cambridge university press, 2019.
  2. R. Vershynin, *High-dimensional probability: An introduction with applications in data science,* Vol. 47. Cambridge university press, 2018.
  3. R. A. DeVore, G. G. Lorents, *Constructive Approximation,* Springer-Verlag, Berlin, 1993.
  4. I. Daubechies, *Ten lectures on wavelets,* CBMS-NSF Regional Conference Series in Applied Mathematics, 61, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.
- Online notes:
  1. F. Bach, *Learning theory from first principles,* MIT press, 2024. Also available at `https://www.di.ens.fr/~fbach/ltfp_book.pdf`.
  2. Arridge S, Maass P, Öktem O, Schönlieb C-B., *Solving inverse problems using data-driven models,* Acta Numerica, 2019;28:1-174, doi:10.1017/S0962492919000059.

**Additional support**

Four examples sheets will be provided and four associated examples classes will be given. There will be a one-hour revision class in the Easter Term. The final exam will consist of six questions, two from each section, with only four questions to be attempted.