Titolo del corso: High-Dimensional Probability for Theoretical Machine Learning

Docente: Dr. Francesco Camilli / Dr. Federica Gerace

Membro del collegio proponente:

Ore frontali di lezione: 20

Periodo di lezione: Gennaio-Febbraio 2027

Settore/i disciplinare del corso: Fisica Matematica

Tipologia di corso: Base

Modalità di verifica dell'apprendimento:

Giornate di Journal Club dove ogni studente presenta un articolo che utilizza come strumenti almeno uno di quelli presentati nel corso.

Abstract del corso:

This PhD course provides an introduction to high-dimensional probability and elements of random matrix theory, with a focus on their applications to contemporary problems in statistics and machine learning. The first part of the course (8 hours) develops the essential analytical tools: concentration of measure in high dimensions and foundational results from random matrix theory. These tools are then applied to the study of high-dimensional inference models, beginning with the spiked Wigner model, where interpolation and replica methods give sharp characterizations of phase transitions.

The second part of the course (10 hours) focuses on applications to statistical learning. We introduce generalized linear models in proportional high-dimensional regimes, covering ridge regression from both random matrix and replica perspectives, and presenting the structure of the proof in the Bayes-optimal setting via Guerra's interpolation. We then examine random feature models, emphasizing their closed-form asymptotics through random matrix theory, replica analyses, and Gaussian Equivalence Principles (GEP), which provide a powerful framework for understanding nonlinear feature maps. Finally, we discuss multilayer perceptrons, presenting recent theoretical results on their high-dimensional behavior and sketching the key ideas behind their analysis, including the proof of GEPs if time allows.

The course concludes (2 hours) with a discussion of significant open problems in high-dimensional statistics and theoretical machine learning, including nonlinear feature learning and limitations of current analytical frameworks. This course is designed for PhD students with interest in the probabilistic foundations of modern inference and machine learning, and the rigorous analysis of high-dimensional models.

Programma del corso:

1. High-dimensional probability, concentration inequalities.

2. Matrix concentration inequalities
3. Interpolation in statistical mechanics and inference: Spiked Wigner model
4. Generalized Linear Models (GLM) e Random Features (RF)
5. Gaussian mixtures
6. Multilayer con GEP in proportional regime
7. Problemi aperti

Timeline indicativa:

- **8hs of introduction** to Concentration inequalities, basics of Random Matrix Theory (RMT), Bayesian inference (Spiked Wigner model (interpolation method/replicas))
- **10hs of applications**: Generalized Linear Models (ridge regression with RMT, proof in the BO setting with interpolation); Random Features (Replicas, closing equations with RMT & Gaussian Equivalence Theorems); Multilayer perceptrons (Replicas with Gaussian Equivalence Principles and their proof)
- **2hs of open problems**