Graduate School Event

Thesis Defense: Trustworthy Machine Learning in High Dimensions

Simone Bombari (Mondelli Group)
Mondelli Group

Host: Matthew Kwan

Artificial intelligence and machine learning have undergone an unprecedented evolution in the past decade, motivating a research effort toward a theory able to capture the qualitative behavior of large-scale neural systems. A central puzzle has been the clear benefit of scaling architecture size and overfitting the training set in supervised learning tasks. This evidence, in apparent contradiction with classical statistical learning theory, pushed researchers to develop a new theory capturing the interplay between the algorithmic and architectural bias of training and the specific target function, differently from previous methods rooted in uniform stability.This approach has enabled a grounded understanding of novel learning regimes, typically through formal limits where the number of training samples n, data dimensions d, and model parameters p grow to infinity at different rates.In this thesis, we follow this approach, focusing on the trustworthiness of high-dimensional models: properties that are difficult to control during training or deployment and often emerge under unpredictable or adversarial conditions. In such settings, it is crucial to formally ensure a priori the reliability of machine learning systems.First, we study data memorization, both as label fitting and as the storage of private information about training samples in trained parameters. We prove that $p = \Omega(n)$ parameters are sufficient for a deep neural network to memorize a generic set of labels, and for a model to memorize spurious features across training data. We then give evidence that $p = \Omega(dn)$ parameters are instead necessary for an adversary to reconstruct the full training set from the trained parameters.Second, we study robustness, both to adversarial perturbations and to distribution shift. We first prove that $p = \Omega(dn)$ parameters can be sufficient for a class of neural networks to overfit the training data while guaranteeing robustness to adversarial perturbations. Then, we focus on spurious correlations learning in high-dimensional regression, studying the effect of the ridge regularization parameter in the proportional regime $n = \Theta(d)$, and connecting it via an equivalence argument to the role of over-parameterization $p = \Omega(n)$ in neural networks. We also investigate the architectural bias of attention-based networks, showing that they are sensitive to the replacement of individual words in an embedded sentence, allowing them to generalize on sentences where the contextual meaning depends on one or few words.Finally, we study differentially private optimization in high-dimensional regimes. We prove that standard private gradient methods do not suffer in the over-parameterized regime $p = \Omega(n)$, challenging the current wisdom based on stability-derived generalization bounds. We then

consider linear regression in the proportional regime $n = \Theta(d)$, showing that standard private gradient descent can achieve optimal rates under appropriate hyper-parameter scaling, such as sufficiently small gradient clipping constants, whose role is still debated in practice.

Wednesday, July 1, 2026 05:00pm - 06:00pm

Office Bldg West / Ground floor / Heinzel Seminar Room (I21.EG.101) and Zoom

---

This invitation is valid as a ticket for the ISTA Shuttle from and to Heiligenstadt Station. Please find a schedule of the ISTA Shuttle on our webpage: https://ista.ac.at/en/campus/how-to-get-here/ The ISTA Shuttle bus is marked ISTA Shuttle (#142) and has the Institute Logo printed on the side.