Generalization in Attention-Based Models: Insights from Solvable High-Dimensional Theory

Lenka Zdeborová

Abstract

Attention layers acting on sequences of tokens are a cornerstone of modern architectures. Understanding generalization in generative AI therefore requires understanding how attention-based systems learn — and generalize — from data. In this talk, I will present recent advances in analyzable, high-dimensional solvable models of learning with attention. Focusing on supervised settings, I will show how these models already yield sharp predictions and useful intuition. I view these results as a stepping stone toward understanding self-supervised and generative training in attention-based models.