Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2026:3**
http://www.math.su.se

Grokking through the lens of information theory

Ashkan Safaee*

May 2026

Abstract

Grokking, the phenomenon where neural networks undergo a sudden transition from memorization to generalization long after the training error has disappeared, remains poorly understood, with implications for AI safety, alignment, and interpretability. Two information theoretic frameworks have been proposed to study representation learning in deep neural networks: Tishby's information bottleneck framework, later adopted by Tishby and Zaslavsky and Shwartz–Ziv and Tishby to deep neural networks, which described an initial fitting phase followed by a representation compression on the information plane, and Clauw et al.'s multivariate framework which identifies feature learning, emergence, and decoupling phases through synergy and redundancy between neurons. This thesis synthesizes the two to build an analytical bridge between them and study the phenomenon of grokking. We propose that the decoupling phase proposed by Clauw et al. corresponds to the representation compression phase predicted in the information bottleneck framework, with weight decay as an enabling factor to drive the network from pure memorization towards a generalizing, compressed representation. Establishing this connection rigorously is left to future work.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ashkansafaee@gmail.com. Supervisor: Ola Hössjer, Leo Levenius.