Relating Loss Function Scaling in Neural Language Models to Hilberg's Conjecture

Adrián Raso González

Vienna University of Technology

IMS International Conference on Statistics and Data Science

Abstract

Hilberg's conjecture posits that block entropy grows sublinearly in natural language data, suggesting a vanishing asymptotic entropy rate and unbounded excess entropy in the limit. This behavior is puzzling and has motivated numerous studies in the literature concerning the theoretical and practical consequences of this growth, having been suggested as an issue for machine learning models. In this work, we ask how this growth interacts with training in neural language models. Using the July 20, 2025 English Wikipedia dump, we estimate block entropies H(n) using per-token code lengths from a bias-corrected Prediction by Partial Matching (PPM) estimator, reporting 95% confidence intervals from variance estimation across 50 disjoint text subsets. We train autoregressive GPT neural models with varied capacity and context length to measure the conditional entropy loss and study how its decrease may relate to the discrete derivative of H(n) and the Hilberg β exponent. The results show that the empirical baseline exhibits subextensive behavior consistent with Hilberg's scaling, and that Transformer LMs display a robust power law improvement of conditional entropy with increasing context, with fitted exponents in a similar subextensive regime.