University of California, Irvine
Statistics Seminar

Decoding Game: On Minimax Optimality of Heuristic Text Generation Strategies

Jason Klusowski
Assistant Professor
Department of Operations Research & Financial Engineering
Princeton University

4 p.m., Thursday, Oct. 17, 2024
6011 Donald Bren Hall

Decoding strategies play a pivotal role in text generation for modern language models, yet a perplexing gap persists between theory and practice. Surprisingly, strategies that should intuitively be optimal, such as Maximum a Posteriori (MAP), often perform poorly in practice. Meanwhile, popular heuristic approaches like Top-k and Nucleus sampling, which employ truncation and renormalization of the conditional next-token probabilities, have achieved great empirical success but lack theoretical justification. This talk introduces the Decoding Game, a theoretical framework that attempts to reconcile this gap by recasting text generation as a two-player zero-sum game. In this game, the Strategist aims to produce text that aligns with the true distribution, while Nature acts as an adversary, distorting the Strategist's target distribution. Our analysis reveals that the adversarial Nature imposes an implicit regularization on the likelihood, with truncation-renormalization methods emerging as first-order approximations of the optimal strategy.