For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics

SEMINAR ON STATISTICS AND DATA SCIENCE

Language Model Watermarking: Detecting and Designing Statistical Dependence

By

Dr. Xiang LI

University of Pennsylvania

Abstract

The widespread use of language models has made the provenance of generated text an increasingly important concern. Major AI providers, including Google, OpenAI, and Anthropic, are adopting text watermarking, while the EU AI Act introduces requirements for machine-readable marking of AI-generated content. Watermarking creates a detectable signal by introducing statistical dependence between generated tokens and keyed pseudorandomness. In the first part, we introduce a statistical framework for detecting and using this dependence. We show how pivotal statistics enable testing without access to the model's token probabilities, develop a minimax perspective on optimal detection, and discuss robustness to human edits. We also briefly describe extensions to estimating the proportion of watermarked content and localizing watermarked passages.

In the second part, we focus on our recent watermark, textGrain, an entropy-calibrated watermark that makes the trade-off between detectability and generation diversity explicit and controllable. Many watermarks preserve the model's next-token distribution when averaged over secret keys, a property called unbiasedness. This guarantee alone, however, does not ensure diversity in repeated generation with the same key, limiting applications that rely on multiple distinct responses. We formulate watermark embedding as an optimal transport problem whose regularization penalty equals the average loss of sampling entropy, linking the dependence used for detection to reduced sampling diversity. This formulation allows us to calibrate the watermark to a prescribed entropy-loss budget. A block-based construction reduces computational cost, preserves unbiasedness, and supports integration with speculative sampling. We compare watermarks at comparable entropy retention and present empirical results on detection power, generation diversity, and task performance.

Bio: Xiang Li is a Postdoctoral Researcher at the University of Pennsylvania and will join the Department of Statistics at Rutgers University as an Assistant Professor in January 2027. He received his Ph.D. in Statistics from Peking University. His research interests include statistics, optimization, and machine learning, with a current focus on the statistical foundations of large language models, particularly watermarking and evaluation.

Date : 9 November 2026 (Monday)
Time : 9:00p.m.-10:00p.m.(Hong Kong Time)
Zoom Meeting ID : 561 696 0008
Passcode : hkust

All are welcome