THE UNIVERSITY OF CHICAGO | DEPARTMENT OF STATISTICS

PhD Dissertation Presentation

Kiho Park
Department of Statistics
The University of Chicago

“The Geometry of Concepts in Large Language Models”

April 23, 2026, at 1:00 PM
Room 103, 5460 S University Ave.

Abstract

This dissertation develops a mathematical foundation for the linear representation hypothesis, the informal idea that high-level semantic concepts are encoded as directions in the representation spaces of large language models (LLMs). Despite growing empirical evidence, a rigorous formulation of what ``linear representation'' means and how the geometry of representation spaces relates to the semantics of language has been lacking. This work addresses this gap through three interrelated studies.

First, we formalize the linear representation of binary concepts and introduce the ``causal inner product'', which connects orthogonality to causal separability. Second, we extend the framework to categorical concepts and show that hierarchical relations are encoded through orthogonality. Third, by identifying the geometry of representation spaces as a Bregman (dually flat) geometry induced by KL divergence, we develop ``dual steering'', a method that modifies a target concept while minimizing off-target interference.

Together, these contributions provide a unified framework for understanding how semantic structure is encoded in the geometry of representations. They offer both theoretical insights into AI interpretability and practical tools for controlled model steering.

For information about building access for persons with disabilities, please contact Keisha Prowoznik at 773.702-0541 or send an email to kprowoznik@uchicago.edu. If you wish to subscribe to our email list, please visit the following website: https://lists.uchicago.edu/web/info/statseminars.