For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics

PHD STUDENT SEMINAR

Sparsity in Linear Attention Models

By

Mr Yuxuan CHEN

Abstract

Linear recurrent architectures, State-Space Models (SSMs), and Linear Attention Large Language Models (LLMs) provide an efficient and compelling alternative to computationally intensive softmax attention, as they compress context into a fixed-size state matrix to enable efficient long-context processing and constant-time inference. However, these models inherently suffer from several core bottlenecks in practical applications: linear recurrent architectures and SSMs are plagued by "stuffed memory" — a phenomenon where fixed-size hidden states become super-saturated, leading to memory collisions and severely degraded retrieval. The research has shown that low-rank heads are indispensable for model reasoning, whereas high-rank heads exhibit significant redundancy. To overcome these optimization bottlenecks, we proposed BregNet, a novel linear sequence architecture grounded in the Linearized Bregman Iteration (LBI) framework, natively induces structural sparsity by decoupling the associative memory into a continuous dual accumulator and a sparse primal state, enforcing a rigorously regularized Inverse Scale Space dynamic governed by l1 and l2 norm through a soft-shrinkage operator, which resolves state saturation while maintaining optimal hardware efficiency. Extensive experiments evaluated across models of varying sizes and on various downstream tasks demonstrate the effectiveness of the aforementioned solutions.

Date : 29 April 2026, Wednesday

Time : 4:00pm

Venue : Room 4472 (Lifts 25/26)

All are Welcome!