UNIVERSITY AT ALBANY
State University of New York

College of Arts and Sciences
Department of Mathematics and Statistics

Analysis and Data Science Seminar

Felix Ye

Entropic Optimal Transport is an Attention

Tuesday, February 10, 2026
3:00 P.M. in Massry B012

Abstract. Entropic optimal transport (EOT) via Sinkhorn iterations is widely used in modern machine learning, yet GPU solvers remain inefficient at scale. Tensorized implementations suffer quadratic HBM traffic from dense $n \times m$ interactions, while existing online backends avoid storing dense matrices but still rely on generic tiled map-reduce reduction kernels with limited fusion. We present **FlashSinkhorn**, an IO-aware EOT solver for squared Euclidean cost that rewrites stabilized log-domain Sinkhorn updates as row-wise LogSumExp reductions of biased dot-product scores, the same normalization as transformer attention. This enables FlashAttention-style fusion and tiling: fused Triton kernels stream tiles through on-chip SRAM and update dual potentials in a single pass, substantially reducing HBM IO per iteration while retaining linear-memory operations. We further provide streaming kernels for transport application, enabling scalable first- and second-order optimization. On A100 GPUs, FlashSinkhorn achieves up to 32× forward-pass and 161× end-to-end speedups over state-of-the-art online baselines on point-cloud OT, improves scalability on OT-based downstream tasks.