Aditya Krishnan

aditya.krishnan94@gmail.com | *+1 412-623-9519* | *Google Scholar* | *linkedin.com/in/arkrishn/*

Research Interests

Similarity search, retrieval augmented generation, scalable machine learning and information retrieval.

Professional Experience

Pinecone Systems, New York City

Senior Research Scientist — *January 2024 — Present*

- Researched improvements to scalability and efficiency (in terms of I/O bandwidth, cache, and compute utilization) of Pinecone's vector index, in particular, designed novel routing mechanism (with publication in submission) for IVF-style vector indexes resulting in 25% reduction in amount of data scanned over state-of-the-art mechanisms like Google's ScaNN.
- Designed novel dimensionality reduction scheme for ANN search using residual networks, beating state of the art mechanisms on popular datasets such as OpenAI Ada02 embeddings of Natural Questions.
- Researching mechanisms to augment LLMs with retrieval for retrieval augmented generation (RAG).

Research Scientist — *October 2022 — January 2024*

- Designed and implemented Pinecone's winning submission to NeurIPS 2023 Big ANN challenge, for the 'Out-of-Distribution' track. Solution in Rust obtains 35K+ QPS on 10M sized index w/ 8vCPUs and 16GB RAM hardware.
- Led research team's effort on implementing quantization of vectors in Pinecone Serverless (Pinecone's multi-tenant vector database offering over blob storage), including SIMD implementations in Rust for low-latency scenarios. Demonstrated a 2x improvement in amount of data cached to serve queries.

Science Intern — *May 2021 — August 2021*

- Researched quantization for similarity search, advised by Edo Liberty (CEO and founder of Pinecone).

Education

Johns Hopkins University, Whiting School of Engineering — *September 2018 — September 2022*

Doctor of Philosophy in Computer Science
Advisor: Vladimir Braverman
Thesis: Fast and Memory-Efficient Algorithms for Matrix Spectrum Approximation

Carnegie Mellon University, School of Computer Science — *May 2017 — May 2018*

Master of Science in Computer Science
Advisor: Anupam Gupta
Thesis: Pricing Online Metric Matching Algorithms on Trees

Carnegie Mellon University, School of Computer Science — *August 2013 — May 2017*

Bachelor of Science in Computer Science with Minor in Engineering Studies

Honors and Awards

JHU MINDS TRIPODS Data Science Fellowship 2022 (awarded to ~5 students across two schools per cycle)
JHU Computer Science Department Fellowship 2018 (awarded to 2 people in incoming class of 50+)
NeurIPS 2022 Top Reviewer (less than 10% reviewers)

Technical Skills

Languages: Rust, Python (Advanced), Java (Beginner)
Libraries: NumPy (Advanced), PyTorch, Distributed Data Parallel, FSDP, SciKit-learn (Intermediate)

Publications

*Authors appear in alphabetical order. Where applicable * denotes equal contribution.*
Sublinear Time Spectral Density Estimation, with Vladimir Braverman and Christopher Musco. *ACM Symposium on Theory of Computing (STOC), 2022.*

Lower Bounds for Pseudo-Deterministic Counting in a Stream, with Vladimir Braverman, Robert Krauthgamer, and Shay Sapir. *International Colloquium on Automata Languages and Programming (ICALP), 2023.*

Lifelong Learning with Sketched Structural Regularization, with Haoran Li, Jingfeng Wu*, Soheil Kolouri*, Praveen K. Pilly and Vladimir Braverman. *Asian Conference on Machine Learning (ACML), 2021.*

Near-Optimal Entrywise Sampling of Numerically Sparse Matrices, with Vladimir Braverman, Robert Krauthgamer, and Shay Sapir. *Conference on Learning Theory (COLT). PMLR, 2021.*

Schatten Norms in Matrix Streams: Hello Sparsity, Goodbye Dimension, with Vladimir Braverman, Robert Krauthgamer, and Roi Sinoff. *International Conference on Machine Learning (ICML), 2020.*

Competitively Pricing Parking in a Tree, with Max Bender, Jacob Gilbert, and Kirk Pruhs. *Conference on Web and Internet Economics (WINE), 2020.*

On Sketching the q to p Norms, with Sidhanth Mohanty and David P. Woodruff. *International Conference on Approximation Algorithms for Combinatorial Optimization Problems (APPROX), 2018.*

Preprints

Optimistic Query Routing for Maximum Inner-Product Search, with Sebastian Bruch and Franco Maria Nardini. 2024. *In Submission.*

Talks

Sublinear Time Spectral Density Estimation, 2022, STOC, Rome, Italy
Sublinear Time Spectral Density Estimation, 2018, JHU CS Theory Seminar, Baltimore
Schatten Norms in Matrix Streams: The Role of Sparsity, 2020, ICML
Schatten Norms in Matrix Streams: The Role of Sparsity, 2019, JHU CS Theory Seminar, Baltimore
Pricing Online Metric Matching Algorithms on Trees, 2018, CMU Theory Seminar, Pittsburgh

Academic Service

Invited Reviewer
NeurIPS 2024, 2023, 2022, 2021
ICML 2024, 2023, 2022, 2021
ICLR 2024, 2023, 2022
STOC 2022, 2021, SODA 2021, PODS 2020
Seminar Organizer
JHU Theory Seminar 2021, 2022
Teaching Assistant
Introduction to Algorithms (JHU) Fall 2019, Spring 2020, Spring 2022
Approximation Algorithms (JHU) Spring 2021

References

Edo Liberty, CEO and Founder, Pinecone, edo@edoliberty.com
Christoper Musco, Assistant Professor, New York University, cmusco@nyu.edu
Vladimir Braverman, Victor E. Cameron Professor, Rice University, vb21@rice.edu