Himadri Mandal

mandalhimadri06@gmail.com • quirtt.github.io • quirtt
quirtt

"If I have seen further it is by standing on the shoulders of Giants" — Isaac A. Newton

Current Interests

Machine Learning, AI Safety, Statistics, Mathematics, Programming, Entrepreneurship, Philosophy

Education

Indian Statistical Institute, Kolkata **August 2023 — present**
Statistics Undergraduate, Semester 1 — 94.4%

Programming experience

Python, Bash, R, Next.js + TailwindCSS, LaTeX, Linux (Arch Linux on i3)

Skills

Leadership, Deductive Reasoning, Debate, Direct Communication, Typing (100+ wpm)

Selected fellowships and awards

ISI K. Semester 1 Outstanding Performance: awarded January 2024, received ₹1500
ISI Kolkata B.Stat. Entrance: awarded August 2023, ranked 11th
IISc Enumeration Finalist: awarded October 2022, ranked top 10
Atlas Fellowship Finalist: awarded September 2022, received 1000$, top 200 in a rationality fellowship
CMI Tessellate Finalist: awarded October 2021
Indian Olympiad Qualifier for Mathematics, KV: awarded February 2021, ranked top 10

Selected Courses

CaMLAB: Cambridge AI Safety Hub *April 8 — April 21*
A course to build the ML engineering fundamentals needed for AI Safety Research including basics of PyTorch, building, training and tuning GPTs and ResNets, interpreting models with TransformerLens, and some emphasis on intro to RL, Deep RL and some RLHF..

Deep Learning: ISI Kolkata *January 2024 — March 2024*
A winter course on deep learning covering Autoencoders, CNNs, GANs, GNNs, Diffusion models, RNNs, Attention mechanics, Transformers, etc.

Measure Theory: Maths Club, ISI Kolkata *December 2023 — February 2024*
Took an introductory course on Measure Theory which helped me understand more of the underpinnings of probability in Statistics.

Last updated June 14, 2024

Linear Algbera: ISI Kolkata *September 2023 — December 2023*

Topped the class with a 99/100. It covered Linear Operators, Matrices, Dual Spaces, Tensor Products, Alternating multilinear maps, Inner Products etc. An extension of it also covered Eigenvalues, Nilpotent Operators, Jordan Canonical Form, Spectral Theorem, Basic Group Representations, etc.

Projects

Research

Circuit Phenomenology Using Sparse Autoencoders: w/ David Udell *Ongoing: June 2024*

Using existing circuit discovery algorithms with sparse autoencoders in open-ended exploration of GPT-2-small. Exploring System 2 thinking, Safety features, Hallucinations etc.

Theoretical

Universal Source Coding: quirtt.github.io/report.pdf

The broad setup is the following: there's data coming in from some source. If the source distribution is known, then Huffman Encoding gives the optimal encoding scheme. This project tries to figure out a good encoding scheme (in multiple contexts!) that guarantees performance against all source distributions!

Independence Is Almost Dependence: quirtt.github.io/posts/nelson/nelson.html

My proof to a theorem: given two independent random variables X, Y you can come up with two new random variables U, V which have the same marginals and ϵ-close joints but are deterministically dependent.

Cold Reflections: quirtt.github.io/

My blog on Mathematics, Statistics, Philosophy and everything else that interests me.

Empirical/Programming

Ponderings on OthelloGPT: quirtt.github.io/posts/OthelloGPT/OthelloGPT.html

MechInterp. OthelloGPT is a GPT model trained on Othello games to predict all the possible legal moves. I look into how the model computes how a certain cell is blank.

ORIGAMI: github.com/quirtt/ORIGAMI/

Implements arXiv:2303.17062, AISTATS 2023. A paper on dimensionality reduction of the support to improve computational efficiency in downstream decision making.

Selected work experience

Software

Website Lead: MTRP, Integration *November 2023 — January 2024*

Deployed a website for the university's fest's annual mathematics competition after learning Next.JS and TailwindCSS, all of it took a week. Maintained it for efficiency and bugfixes. See mtrp.integrationfest.in. The github repo: github.com/Integration-ISIK/mtrp-2024-website

Volunteer and outreach

Owner: awas *October 2020 — October 2022*

Served as the **organizer and mentor** for daily math problem solving sessions, philosopical debates, and programming discussions for over two years on **Discord**. Mentored smart math enthusiasts, from all over India, learn hard math, and guided a few of them to get into the Indian training camp of IMO.

Last updated June 14, 2024