

Swagato Das

 swagato.vercel.app
 github.com/Swag-Pseudopy
 swagato.isi2227@gmail.com
 [Swagato Das](#)
 [Swagato Das](#)

EDUCATION

Indian Statistical Institute, Kolkata <i>Masters of Statistics[M.Stat.] - 1st Semester</i>	July 2025 - Present <i>Aggregate Score - Ongoing</i>
Indian Statistical Institute, Kolkata <i>Bachelors of Statistics[B.Stat.](Honors)</i>	August 2022 - May 2025 <i>Aggregate Score - 72%</i>
Hem Sheela Model School, Durgapur AISSCE <i>All India Senior Secondary Certificate Examination</i>	2021 <i>Aggregate Score - 94.2%</i>
Hem Sheela Model School, Durgapur AISSE <i>All India Secondary School Examination</i>	2019 <i>Aggregate Score - 95%</i>

RELEVANT COURSEWORK

Courses: Causal Inference, Statistical Methods of Genetics, Martingale Theory, Functional Analysis, Bayesian Inference, Time Series Method, Statistical Computing, Pattern Recognition, Sample Surveys, Design of Experiments, Parametric Inference, Nonparametric and Sequential Methods, Large Sample Statistical Methods, Linear Statistical Models, Decision Theory, Multivariate Analysis, Regression Techniques, Categorical Data Analysis, Molecular Biology(Elective), Introduction to Stochastic Processes, Applied Stochastic Processes, Introduction to Programming and Data Structures, Real Analysis (I-III), Probability Theory (I-III), Measure Theoretic Probability, Vector and Matrices (I-II), Numerical Analysis, Elements of Algebraic Structures, Discrete Mathematics, Differential Equation, Design and Analysis of Algorithms, Statistical Quality Control and Operations Research, Statistics Comprehensive, Metric Topology and Complex Analysis, Optimization Techniques.

EXPERIENCE

Intelligent Shopfloor <i>Intern</i>	May 2026 - July 2026
- In this project, we aim towards developing an intelligent pipeline for the mechanical industry utilizing neuro-symbolic AI to generate prescriptive operational actions rather than standard diagnostic alerts. - Implementing causal inferential logic to accurately map the directionality component of Directed Acyclic Graphs (DAGs) derived from complex, domain-specific knowledge graphs. - Bridging symbolic reasoning and machine learning to facilitate robust causal discovery, allowing the AI to understand and act upon systemic dependencies within plant environments.	
TU Darmstadt, Self Organizing Systems Lab <i>Research Intern</i>	May 2025 - July 2025
- In this project, we worked on the development of a deep learning system for in-silico SELEX simulation and prediction. - Designed and implemented a cross-attention predictor to model round-wise DNA sequence enrichment. - Building a generative framework using flow matching and neural ODEs to simulate the full enrichment trajectory. - Responsible for constructing embedding pipelines and training conditional vector fields. - Evaluating models for predictive accuracy and generative performance. - Work integrates bioinformatics, generative modeling, and dynamic system learning.	

- As a member of the Organizing Committee for the Winter School on Deep Learning 2024-26 (WSDL 2024-26), I contributed to an intensive academic event aimed at providing self-motivated students with in-depth knowledge and exposure to cutting-edge research in Deep Learning.
- Organized by the Electronics and Communication Sciences Unit at the Indian Statistical Institute, Kolkata, the school featured hands-on courses focused on applying state-of-the-art deep learning techniques to real-world problems.
- It also covered essential theoretical foundations, including core mathematical concepts and programming skills, fostering a comprehensive understanding of the field.

- A small organization focusing on promoting interest and understanding in mathematics beyond the formal classroom setting, run and managed by mathematics enthusiasts at Indian Statistical Institute, Kolkata

PUBLICATIONS, PRE-PRINTS AND UNDER REVIEW ARTICLES

Hyperbolic Fuzzy C-Means with Adaptive Weight-based Filtering for Clustering in Non-Euclidean Spaces | *Accepted at the ICVGIP, 2025 in Oral and Poster category*

[scholar][arxiv]

- **Authors :** Das S., Pratihari A., Das S.
- **Overview :** In this paper we introduce Filtration-based Hyperbolic Fuzzy -Means (HypeFCM), a novel clustering algorithm tailored for better representation of data relationships in non-Euclidean spaces. HypeFCM integrates the principles of fuzzy clustering with hyperbolic geometry and employs a weight-based filtering mechanism to improve performance. The algorithm initializes weights using a Dirichlet distribution and iteratively refines cluster centroids and membership assignments based on a hyperbolic metric in the Poincaré Disc model. Extensive experimental evaluations demonstrate that HypeFCM significantly outperforms conventional fuzzy clustering methods in non-Euclidean settings, underscoring its robustness and effectiveness.

INTERESTS

Biostatistics, Causal Inference and Discovery, Bayesian Techniques, Statistical Learning, Non-Parametric Clustering techniques, Flow Matching, Diffusion Models, LLMs, Deep Learning, Machine Learning, Convex and Non-Convex Optimization Techniques, Convex Clustering, Evolutionary Algorithms, Reinforcement Learning, Evolutionary Reinforcement Learning, Information Theory, Graph Learning.

SKILLS

Languages: Python, R, C, $\text{\LaTeX}$, Java, Julia

Tools: Linux/Bash, RStudio, Git/GitHub, Claude Code, WandB, VS Code, Google AppScripts, Jupyter Notebook, Google Colab, Microsoft Copilot, ChatGPT, Geogebra

PROJECTS

Posterior Sampling from a Network via Rectified Flow | *ISI, Kolkata*

May, 2025 - Present

- **Guide :** Dr. Sagnik Nandy [Ohio State University], Dr. Soham Bonnerjee [University of Pennsylvania]
- **Overview :** This work addresses Bayesian parameter estimation in network models incorporating both graph topology and node-level covariates. We propose a novel framework where the prior is implicitly defined through a parameterized transport mapping from a base noise distribution to a feature-conditioned target. To navigate the intractable posterior, we introduce a family of sampling algorithms that seamlessly integrate the complex implicit prior with the network likelihood. Our primary contribution includes both an exact iterative sampling scheme and a highly scalable approximate variant that relies on a closed-form analytical surrogate to bypass expensive numerical integration steps. Theoretical justifications for the proposed inference mechanisms are provided, alongside extensions for sparse network regimes.

Investigating Tree of Thoughts Prompting for Mathematical Reasoning in Large Language Models | *ISI, Kolkata*

December, 2025 - February, 2026

- **Guide** : Dr. Promit Ghoshal [University of Chicago]
- **Overview** : In this short-term exploratory project, we investigate the effectiveness of the Tree of Thoughts (ToT) prompting framework in enabling Large Language Models (LLMs) to solve simple mathematical reasoning problems more reliably than standard direct prompting. The project studies BFS- and DFS-based search strategies where the LLM generates and evaluates intermediate reasoning steps (“thoughts”) while navigating a structured search tree. Experiments are conducted on tasks such as the Game of 24 and elementary linear equation solving, with emphasis on reasoning consistency, search-guided problem solving, and interpretability of intermediate thought processes.

Escaping Preference Collapse: Cycle Dynamics in Nash Learning from Human Feedback | *ISI, Kolkata*

May, 2025 - May, 2026

- **Guide** : Prof. Swagatam Das
- **Overview** : Investigated Nash Learning from Human Feedback (NLHF) as a two-player zero-sum game to resolve non-transitive Condorcet cycles in LLM alignment. Conducted a rigorous dynamical-systems analysis, deriving a universal constant of motion for replicator dynamics and establishing topological bounds on policy orbits. Proved a strict Stability Bifurcation governed by the τ/η ratio in discrete Nash mirror-descent updates, extending theoretical guarantees to stochastic settings and symmetric k -action cycles. Validated these stability regimes empirically across synthetic tournaments, parameterized neural networks, and a 124M-parameter GPT-2 initialization.

Intersection of deterministic curves | *ISI, Kolkata*

May, 2025 - Present

- **Guide** : Dr. Wasiur Rahman Khuda Buksh [University of Nottingham] and Dr. Olga Iziuntseva [University of Nottingham]
- **Overview** : In this project, we investigate the asymptotic behavior of self-intersection integrals for deterministic curves to establish rigorous scaling bounds. We initially ground our analysis in polynomial traversals, leveraging the density of monomials within the space of continuously differentiable functions with compact support. However, extending this framework reveals critical analytical hurdles, specifically a discrepancy between the upper and lower bounding rates caused by inherent traversal pathologies. To address these challenges, we introduce and define a novel “regularized class” of functions. This regularization successfully resolves the identified bounding discrepancies, allowing us to formally establish matching upper and lower bounds that share the exact same stable asymptotic convergence rates.

Analysis of Low-Pass and High-Pass Filters on Graph Features | *ISI, Kolkata*

May, 2025 - August, 2025

- **Guide** : Dr. Kushal Bose, Prof. Swagatam Das
- **Overview** : In this study, we investigate the effects of low-pass and high-pass filters, represented as $(I - L)^k$ and $(I + L)^k$, where, L denotes the graph laplacian, on the features of heterophilic and homophilic graphs. Low-pass filters are commonly used in signal processing and graph analysis to emphasize the smooth, low-frequency components of the data, whereas high-pass filters highlight the high-frequency components, effectively suppressing the smooth variations. We apply these filters to datasets characterized by either homophilic or heterophilic structures. In the case of homophilic graphs, vertices that share similar properties tend to be more strongly connected. In contrast, heterophilic graphs have edges between nodes with dissimilar properties. By analyzing how these filters affect the feature distributions in both graph types, we seek to derive insights into the graph structure and how transformations like these might aid in tasks such as classification, anomaly detection, or community detection.

Antithetic Noise in Diffusion Models | *ISI, Kolkata*

March, 2026 - Present

- **Guide** : Dr. Yuchen Wu [Cornell University]
- **Overview** : In this project, motivated from the recent ICLR 2026 paper Antithetic Noise in Diffusion Models, by Jia et al, we tried to understand how the authors describe the effect of antithetic sampling in Diffusion Models, which supposedly is a technique well prevalent in the literature of Variance reduction and try to interpret the concepts defined in the more general sampling frameworks.

A Generalized Framework for Penalized Local Mode-Seeking: Unifying Mean Shift and Sparse Smoothing Splines | *ISI, Kolkata*

November, 2025 - December, 2025

- **Guide** : Prof. Swagatam Das
- **Overview** : In this project we tried to look at the mode-seeking algorithms, such as the classic Mean Shift, which are foundational tools for non-parametric clustering and density analysis. However, standard formulations rely on zero-degree local constant estimators, which suffer from boundary bias and oversmoothing of high-curvature density peaks. In this project, we propose a generalized umbrella framework for mode-seeking, formulated as a Penalized Local Functional Maximization problem. We demonstrate that existing mode-seeking algorithms are special cases of this framework. Furthermore, we introduce a novel instantiation of this framework: the Sparse Smoothing Spline Mean Shift (SSS-MS). By employing a dual-penalty structure—an L_2 integral roughness penalty and an L_1 Group Lasso complexity penalty—the SSS-MS adaptively selects its own functional complexity at every spatial location, ensuring convergence properties that theoretically strictly dominate or equal the performance of its standard generalizations.

Bridging the Gap in Constrained Online Convex Optimization via the Polyak-Łojasiewicz Condition | *ISI, Kolkata*

November, 2025 - May, 2026

- **Guide** : Prof. Swagatam Das jointly with Dr. Abhishek Sinha [TIFR, Mumbai]
- **Overview** : In this project we tried extending the optimal bounds in Constrained Online Convex Optimization (COCO) to overparameterized ML models by utilizing the Polyak-Łojasiewicz (PL) condition. Proved that a novel drift-plus-penalty framework, proposed by Sinha et. al. successfully bypasses standard $\mathcal{O}(\sqrt{T})$ limits to achieve $\mathcal{O}(\log T)$ static regret and $\mathcal{O}(\sqrt{T \log T})$ cumulative constraint violations

Resampling Methods for Variance Estimation in G-Computation: A Comparative Study under Selection Bias | *ISI, Kolkata*

May, 2025 - October, 2025

- **Guide** : Dr. Jayant Jha
- **Overview** : In this project, we investigated variance estimation for g-computation under complex selection bias by replacing traditional, analytically derived M-estimation frameworks (Zivich et al., 2026) with three resampling techniques: non-parametric bootstrap, m-out-of-n bootstrap, and delete-d jackknife. Large-scale Monte Carlo simulations proved that these finite-sample resampling methods match the asymptotic sandwich estimator's performance, achieving nominal 95% coverage and calibrated standard errors. This approach trades computational speed for significant improvements in implementation simplicity.

Statistical Modelling of Prostate Cancer Biomarkers: A Comparative Regression Analysis | *ISI, Kolkata*

November, 2025

- **Guide** : Dr. Abhik Ghosh
- **Overview** : In this project, we did a statistical analysis of prostate cancer data, aiming to model the Log Prostate Specific Antigen (lpsa) using various clinical covariates. Utilizing a dataset of 97 patients, we employed Multiple Linear Regression, Robust Regression, Lasso Regularization, and Generalized Additive Models (GAM). Our findings suggested that despite the presence of influential observations, linear structures dominate the data generating process, with Lasso Regression providing the optimal balance between bias and variance (Test RMSE: 0.7037). The study highlighted the critical role of tumor volume and capsular penetration in predicting PSA levels.

Investigating Virtual Evolution: SELEX Simulator with Generative Models | *TU Darmstadt, Germany*

May, 2025 - October, 2025

- **Guide** : Prof. Dr. Heinz Koepl, Philipp Froehlich, Sebastian Wirth
- **Overview** : In this project, we worked on the development of a deep learning system for in-silico SELEX simulation and prediction. Designed and implemented a cross-attention predictor to model round-wise DNA sequence enrichment, and currently building a generative framework using flow matching and neural ODEs to simulate the full enrichment trajectory. Responsibilities include constructing embedding pipelines, training conditional vector fields, and evaluating models for predictive accuracy and generative performance. The work integrates bioinformatics, generative modeling, and dynamic system learning.

January, 2025 - May, 2025

- **Guide** : Prof. Swagatam Das jointly with Dr. Saptarshi Chakraborty [University of Michigan] and Debolina Paul [PhD, Oxford University]
- **Overview** : In this project, We ventured through and explored various techniques that are prevalent in the landscape of the Convex clustering literature and various methodologies that commonly are used by practitioners to deal with the sort of optimization problems these usually transform into, namely ADMM [Alternating Direction Method of Multipliers], and using these we were attempting to adapt to situations where clusters w.r.t. both the rows and columns are of substance, a.k.a. biclustering, hence, devising an appropriate objective function(s) coupled with appropriate optimization subroutines to reach to meaningful clusters, for instance, clustering a group of samples and along with that a group of genes simultaneously and so on.

Various Regression methodologies | ISI, Kolkata

August, 2024 - May, 2025

- **Guide** : Prof. Kiranmoy Das
- **Overview** : In this project, We have explored and assessed various kinds of penalizations prevalent in the literature of regression and how they affect the effectiveness of the model in terms of it's ability in tactfully handling certain commonly observable traits in real life data and otherwise like sparsity, Multicollinearity and Model Misspecification.

Various Bregman Divergences and clustering based on those divergences and their consistency and convergence rates under different convergence subroutines | ISI, Kolkata

October, 2024 - February, 2025

- **Guide** : Prof. Swagatam Das
- **Overview** : In this project, We have explored various Bregman divergences such as the Kullback Leibler divergence, euclidean distance and so on, and studied how they affect the clustering ability of an algorithm and how they are able to identify hidden structures or patterns in data without any human intervention or supervision. We also tried to analyse their convergence rates under various optimization subroutines like Gradient descent, ADAM and so on and hence obtain some asymptotic bounds on these rates along with which we also explored various convex formulations of such objectives and tried to observe how they influence each other.

On Reinforcement Learning, Evolutionary algorithms | ISI, Kolkata

January, 2024 - September, 2024

- **Guide** : Prof. Swagatam Das
- **Overview** : This Project is mostly about surveying literature in these disciplines and how these disciplines have intertwined and gave rise to the techniques commonly used. In this project, I implemented algorithms like minimax, Q-learning, MCTS and a few other variants, especially the evolutionary variants of it including a few inspired from deep learning methodologies on a few simple gym environments like tic tac toe, blackjack, ludo and a few more atari environments.

On the performance of PLUG-IN Method and LEAVE-ONE-OUT Method in construction of an estimate of the intensity parameter | ISI, Kolkata

April, 2024

- **Guide** : Prof. Probal Chaudhuri
- **Overview** : This class project mainly focused on a comparative study of the PLUG-IN and the LEAVE-ONE-OUT Methods which are used quite often in estimation of the intensity parameter where, we want to obtain an estimate for $\mu(\cdot)$ from the sample at hand, starting with a consistent estimate of the density(given, it exists), we construct a point-wise local average estimate of the parameter by considering a δ -neighborhood($V_\delta(t)$ or maybe, a deleted one $V'_\delta(t)$) around the desired t , say $t_0 \in [0, 1]$, hence the estimate for $\mu(t_0)$, namely $\hat{\mu}(t_0)$ [more appropriately, $\hat{\mu}_\delta(t_0)$]; these methods intervene to determine the optimal δ [the size of the moving window] for which the MSE of the estimate is in a reasonable vicinity of the Global minima(if exists) of the MSE over all the δ s.

Exploring Spectral Clustering, Kernel based methods for clustering like kernel power K-means and Dirichlet Process Mixture Models based clustering approaches and it's variants | ISI, Kolkata

August, 2023 - December, 2023

- **Guide** : Prof. Swagatam Das
- **Overview** : This project is primarily centered around studying algorithms like K-means, Power K-means, Kernel K-means, Gaussian Mixture Models, where while initializing we need to state the number of clusters, and developing it into one where it itself determines the required number of clusters using Dirichlet Process Mixture Model like approach for optimal Loss, detects a broader variety of cluster instead of only the "convex" ones like in Spectral clustering like algorithms and is less sensitive to the hyperparameters.

Reading Project on Spectral Clustering and it's variants | *ISI, Kolkata*

May, 2023 - September, 2023

- **Guide** : Prof. Malay Bhattacharyya
- **Overview** : This project is mainly focused on studying and reproducing the traditional Spectral Clustering Algorithm as was suggested by Andrew Y. Ng, Micheal I. Jordan and Yair Weiss in their paper published on 2001 "*On Spectral Clustering: Analysis and an algorithm.*" and a few variants of it namely, the one stated in the paper "*Eigen Selection in Spectral Clustering: A theory-Guided Practice*" by Xiao Han, Xin Tong and Yingying Fan(accepted on April 2021).

Body Performance Analysis | *ISI, Kolkata*

May, 2023

- **Guide** : Prof. Kiranmoy Das
- **Overview** : This project is all about exploratory data analysis using multiple linear regression, univariate quantile regression and simulations of multiple variables for prediction of missing values on body performance data from a Korean sports promotion foundation.

Reading Project on Quantile Regression and Its Applications | *ISI, Kolkata*

February, 2023

- **Guide** : Prof. Kiranmoy Das
- **Overview** : This project is centered around study and comparison of exposure groups on the distribution of an outcome

Bivariate Analysis of Nitrogen Use Efficiency and Nitrogen Fertilizer Use per hectare of cropland for India and Cuba between 1991-2014 | *ISI, Kolkata*

December, 2022

- **Guide** : Prof. Kiranmoy Das
- **Overview** : This project is an exploratory analysis of Nitrogen Use Efficiency and Nitrogen Fertilizer Use per hectare of cropland and a quest for whether it is possible to cut down on the amount nitrogenous fertilisers used in agriculture without affecting the yield.
- This analysis provides us an assessment of the Cuban system of Organopónicos, and a vision for moving Indian agriculture towards sustainability.

Exploratory Data Analysis on Life Expectancy and GDP per capita of Russia,1951-2022 | *ISI, Kolkata*

November, 2022

- **Guide** : Prof. Kiranmoy Das
- **Overview** : This project is a preliminary analysis of how the life expectancy (in years) and the GDP per capita (in international dollars, fixed 2017 prices, PPP based on 2017 ICP) has been varying over the past 7 decades in Russia.

ACADEMIC ACHIEVEMENTS

Indian Statistical Institute

Qualified for direct admission to M.Stat. program with a government funded stipend.

Indian Statistical Institute

Admitted for B.Stat.(Hons.) program with a government funded stipend.

Joint Entrance Examination Mains

Ranked amongst top 0.1 percentile in Joint Entrance Examination - Mains amongst 1 Million candidates.

LANGUAGES

English : Full professional proficiency

Bengali : Native or bilingual proficiency

Hindi : Native or bilingual proficiency

HOBBIES

Music, Table Tennis, Coding