Pragya Sur

Curriculum Vitae

Department of Statistics Harvard University 1 Oxford Street, SC 712, Cambridge, MA-02138 🖸 pragya@fas.harvard.edu http://sites.fas.harvard.edu/~prs499/index.html

Employment

- July, 2020 Assistant Professor, Department of Statistics, Harvard University
 - 2019-2020 **Postdoctoral Fellow**, Center for Research on Computation and Society, John A. Paulson School of Engineering and Applied Sciences, Harvard University
 - 2017 Summer Research Intern, Microsoft Research

Education

- 2014-2019 Ph.D., Department of Statistics, Stanford University
 Thesis title: A Modern Maximum-Likelihood Theory for High-Dimensional Logistic Regression (Winner of Theodore W. Anderson Theory of Statistics Dissertation Award)
- 2012-2014 Master of Statistics, Indian Statistical Institute, Kolkata, India
- 2009-2012 Bachelor of Statistics (Hons.), Indian Statistical Institute, Kolkata, India

Awards and Honors

- 2024 Supervised the **Harvard Hoopes Prize** winning thesis "*High-Dimensional Linear Interpolation for Structured Data*" by Kevin Luo.
- 2024 Invited to serve as Associate Editor for Statistical Science.
- 2024 Invited visitor at the **Cowles Foundation for Research in Economics**, Yale University.
- 2024 Awarded the Eric and Wendy Schmidt Fund for Strategic Innovation.
- 2023 International Strategy Forum (ISF) Fellow, an 11-month, non-residential fellowship program for rising leaders aged 25 35 from across Africa, Asia, North America, and Europe.
- 2023-2024 Awarded the Dean's Competitive Fund for Promising Scholarship.
- 2022-2024 President of the Institute of Mathematical Statistics (IMS) New Researchers Group (IMS Bulletin article on my contributions: https: //imstat.org/2023/02/15/new-researchers-group-news-2/; Internal Harvard article on my leadership vision and the cohesive link between my roles at IMS and as an ISF Fellow: https://statistics.fas.harvard.edu/news/ professor-pragya-sur-nrg-president-and-isf-2023-asia-fellow).
- 2022, 2024 Invited panelist for the National Science Foundation.

- 2021 Invited speaker at the National Academies of Sciences, Engineering, and Medicine Symposium on "Mathematical Challenges in Machine Learning and Artificial Intelligence" (https://www.nationalacademies.org/event/11 -11-2021/mathematical-challenges-for-machine-learning-and-artific ial-intelligence).
- 2021-2025 National Science Foundation, Division of Mathematical Sciences, Award No. 2113426.
 - 2021 Invited Long-Term Participant for the program "Computational Complexity of Statistical Inference", Simons Institute for the Theory of Computing, University of California, Berkeley.
- 2020-2022 William F. Milton Fund Award, Harvard University.
 - 2019 Theodore W. Anderson Theory of Statistics Dissertation Award, Stanford University.
- 2017-2019 Ric Weiland Graduate Fellowship in the Humanities and Sciences, Stanford University.
 - 2017 Summer Research Internship, Microsoft Research.
 - 2016 Differential Privacy Travel Award, Sloan Foundation.
 - 2013 Usri Gangopadhyay Memorial Gold Medal, for academic performance in the Bachelor of Statistics program, Indian Statistical Institute.
 - 2013 Mukul Chaudhuri Memorial Award, for the highest scoring female student in Bachelor of Statistics (2010-2011), Indian Statistical Institute.
 - 2012 Selected as the inaugural candidate for Summer Exchange Program between Department of Biostatistics, Johns Hopkins University, and Indian Statistical Institute.
 - 2011 Indian Academy of Sciences Summer Research Fellowship.
 - 2011 Jawaharlal Nehru Centre for Advanced Scientific Research Summer Research Fellowship.
- 2010-2013 Kishore Vaigyanik Protsahan Yojana or KVPY (English: Young Scientist Incentive Plan) Fellowship, Department of Science and Technology, Government of India.
 - 2010 Innovation in Science Pursuit for Inspired Research Scholarship, Department of Science and Technology, Government of India.

Research Interests

- High-dimensional Statistics
- o Machine Learning and Artificial Intelligence
- Learning under Heterogeneity and Distribution Shifts
- Causal Inference in High Dimensions

Preprints and Publications 2024+ Generalization Error of Min-Norm Interpolators in Transfer Learning -with Yanke Song and Sohom Bhattacharya, arXiv identifier: https://arxiv.org/pd f/2406.13944, In Review at the Annals of Statistics. 2024+ HEDE: Heritability Estimation in High Dimensions by Ensembling **Debiased Estimators** -with Yanke Song and Xihong Lin, arXiv identifier: https://arxiv.org/abs/2406.1 1184, In Review at Journal of the American Statistical Association. 2024+ ROTI-GCV: Generalized Cross-Validation for right-ROTationally Invariant Data -with Kevin Luo and Yufan Li, arXiv identifier: https://arxiv.org/abs/2406.11666, In Review at AISTATS. 2024+ Predictive Inference in Multi-environment Scenarios -with Kuanhao Jiang, John Duchi, Suyash Gupta, arXiv identifier: https://arxiv.or g/abs/2403.16336, Statistical Science (minor revision). 2024 Universality in Block Dependent Linear Models with Applications to **Non-Parametric Regression** -with Samriddha Lahiry, IEEE Transactions on Information Theory doi: 10.1109/TIT.2024.3481956. 2023+ Spectrum-Aware Debiasing: a Modern Inference Framework with **Applications to Principal Components Regression** -with Yufan Li, arXiv identifier: https://arxiv.org/pdf/2309.07810.pdf, In Review at the Annals of Statistics. 2023 High-dimensional Asymptotics of Langevin Dynamics in Spiked Matrix Models -with Tengyuan Liang and Subhabrata Sen, Information and Inference: A Journal of the IMA, Volume 12, Issue 4, December 2023, Pages 2720–2752. 2022+ A New Central Limit Theorem for the Augmented IPW Estimator: Variance Inflation, Cross-fit Covariance and Beyond -with Kuanhao Jiang, Rajarshi Mukherjee and Subhabrata Sen, arXiv identifier: https: //arxiv.org/abs/2205.10198, The Annals of Statistics (Minor Revision). (Jiang received the 2022 New England Statistical Society Student Research Award for this work.) 2022+ Multi-study Boosting: Theoretical Considerations for Merging versus Ensembling -with Cathy Shyr, Giovanni Parmigiani and Prasad Patil, arXiv identifier: https: //arxiv.org/abs/2207.04588, In Review at Journal of Machine Learning Research. 2022 A Non-Asymptotic Moreau Envelope Theory for High-Dimensional **Generalized Linear Models** -with Lijia Zhou, Frederic Koehler, Danica Sutherland and Nathan Srebro, NeurIPS. 2022 A Precise High-Dimensional Asymptotic Theory for Boosting and Minimum-L1-Norm Interpolated Classifiers -with Tengyuan Liang, The Annals of Statistics, 50.3 (2022): 1669-1695. 3

- 2022 The Asymptotic Distribution of the MLE in High-dimensional Logistic Models: Arbitrary Covariance
 -with Qian Zhao and Emmanuel Candès, Bernoulli 28(3): 1835-1861 (August 2022). DOI: 10.3150/21-BEJ1401.
- 2020 Representation via Representions: Domain Generalization via Adversarially Learned Invariant Representation
 -with Zhun Deng, Frances Ding, Cynthia Dwork, Rachel Hong, Giovanni Parmigiani and Prasad Patil, arXiv identifier: https://arxiv.org/pdf/2006.11478.pdf.
- 2020 Abstracting Fairness: Oracles, Metrics, and Interpretability -with Cynthia Dwork, Christina Ilvento and Guy Rothblum, Proceedings of the Symposium on the Foundations of Responsible Computing, volume LIPIcs, Volume 156, FORC 2020.
- 2020 The Phase Transition for the Existence of the Maximum Likelihood Estimate in High-Dimensional Logistic Regression -with Emmanuel Candès, The Annals of Statistics 48, (2020), no. 1, 27-42.
- 2019 A Modern Maximum-Likelihood Theory for High-Dimensional Logistic Regression
 -with Emmanuel Candès, Proceedings of the National Academy of Sciences 116 (2019), no. 29, 14516–14525.
- 2019 The Likelihood Ratio Test in High-Dimensional Logistic Regression Is Asymptotically a Rescaled Chi-Square
 -with Yuxin Chen and Emmanuel Candès, Probability Theory and Related Fields 175 (2019), no. 1-2, 487–558.
- 2015 Modeling Bimodal Discrete Data using Conway-Maxwell-Poisson Mixture Models

-with Galit Shmueli, Smarajit Bose and Paromita Dubey, Journal of Business and Economic Statistics, 33, no. 3: 352- 365.

2013 Fitting Com-Poisson Mixtures to Bimodal Count Data -with Smarajit Bose, Galit Shmueli and Paromita Dubey, Proceedings of the 2013 International Conference on Information, Operations Management and Statistics (winner of Best Paper Award at ICIOMS 2013).

Invited Talks

- 2025 Workshop on "Bayesian Uncertainty Quantification in Large Models", Banff International Research Station, Chennai, India
- 2025 JSM, Nashville, Tennessee
- 2025 IMS Special Invited Session, ENAR Spring Meeting, New Orleans, LA
- 2024 International Indian Statistical Association Conference, Cochin, India
- 2024 IMS International Conference on Statistics and Data Science, Nice, France
- 2024 CREST-CMAP Statistics Seminar, ENSAE, France
- 2024 **Econometrics Seminar**, Cowles Foundation for Research in Economics, Yale University

- 2024 Conference on Statistical Learning and Data Science, Newport Beach, California
- 2024 Big Data and Machine Learning in Econometrics, Finance, and Statistics, UChicago
- 2024 Statistics Colloquium, Washington University of St. Louis
- 2024 Econometrics Seminar, NYU
- 2024 Statistics Colloquium, Columbia University
- 2024 Bernoulli-IMS 11th World Congress in Probability and Statistics, Bochum, Germany
- 2024 Workshop on High-Dimensional Learning Dynamics: The Emergence of Structure and Reasoning, *ICML*, *Vienna*
- 2024 International Symposium on Nonparametric Statistics, Braga, Portugal
- 2024 Workshop on Statistical Network Analysis and Beyond, Nassau, Bahamas
- 2024 **Statistical Machine Learning for High-Dimensional Data**, Institute of Mathematical Sciences, National University of Singapore (presented online due to visa constraints)
- 2024 Statistics Colloquium, Rutgers University
- 2024 58th Annual Conference on Information Sciences and Systems, Princeton University
- 2023 Big Data and Machine Learning in Econometrics, Finance, and Statistics, University of Chicago
- 2023 Inference and Discovery in Complex Large High-dimensional Data, Invited Paper Session, Joint Statistical Meetings
- 2023 **Statistical Physics and Machine Learning Back Together Again**, Institut d'Etudes Scientifiques de Cargèse, Corsica
- 2023 Learning and Inference from Structured Data: Universality, Correlations and Beyond, The Abdus Salam International Centre for Theoretical Physics (ICTP), Italy
- 2023 FLAIR Seminar Series, EPFL
- 2023 Approximation Methods in Bayesian Analysis, CIRM Luminy
- 2023 Graphical Models, Statistical Inference and Algorithms, Center of Mathematical Sciences and Applications, Harvard University
- 2023 Recent Advances in Statistics and Data Science (Profs. Regina Liu and Cun-Hui Zhang Birthday Conference), *Rutgers University*
- 2023 Statistics Seminar, Cornell University
- 2023 MIT/Harvard Econometrics Seminar, Harvard University
- 2023 Statistical Machine Learning Workshop, Columbia University
- 2023 Lightning Talk, International Strategy Forum Global Summit, Schmidt Futures, Los Angeles

- 2022 International Indian Statistical Association Conference, Indian Institute of Science, Bengaluru
- 2022 IMS International Conference on Statistics and Data Science, Florence, Italy
- 2022 Neyman Seminar, Department of Statistics, University of California, Berkeley
- 2022 Statistics and Probability Seminars, School of Mathematical Sciences, University of Nottingham, online due to Covid-19
- 2022 Big Data and Machine Learning in Econometrics, Finance, and Statistics, University of Chicago
- 2022 Colloquium Seminar, Toyota Technological Institute at Chicago
- 2022 Deep Learning Theory Workshop and Summer School, Simons Institute for the Theory of Computing, University of California, Berkeley
- 2022 Group Meeting for the Yu Group, University of California, Berkeley, online due to Covid-19
- 2022 Joint Statistics and Econometrics Seminar, University of California, Riverside, online due to Covid-19
- 2022 International Seminar on Selective Inference, Online Seminar Series
- 2022 Statistics and Data Science Conference, Keynote Speaker, Massachusetts Institute of Technology (MIT)
- 2022 56th Annual Conference on Information Sciences and Systems, Princeton University, online due to Covid-19
- 2022 Workshop on "Hot Topics: Foundations of Stable, Generalizable and Transferable Statistical Learning", Mathematical Sciences Research Institute, Berkeley, online due to Covid-19
- 2021 Stochastics and Statistics Seminar, MIT
- 2021 Mathematical Challenges in Machine Learning and Artificial Intelligence, Symposium of National Academies' Board on Mathematical Sciences and Analytics, online due to Covid-19
- 2021 Probability Seminar, University of Waterloo, online due to Covid-19
- 2021 Statistics Seminar, University of Toronto, online due to Covid-19
- 2021 **Seminar**, Johns Hopkins Mathematical Institute for Data Science, online due to Covid-19
- 2021 Weekly Seminar, Computational Complexity of Statistical Inference program, Simons Institute for the Theory of Computing, UC Berkeley
- 2021 Online session on "Mean field asymptotics in high-dimensional statistics", related to the 12th International Conference on Multiple Comparison Procedures
- 2021 **Statistical Learning Seminars**, Online seminar series co-organized by The Department of Mathematics, Wroclaw University, The Department of Mathematics, University of Burgundy, and The Department of Statistics, Lund University
- 2021 Statistics Colloquium, Texas A & M University, online due to Covid-19

- 2021 Discussant Talk, International Seminar on Selective Inference
- 2021 Seminar, Statistical Laboratory, University of Cambridge, online due to Covid-19
- 2021 Centre for Research in Statistical Methodology Seminar Series, Department of Statistics, The University of Warwick, online due to Covid-19
- 2020 Applied Math Colloquium, Brown University, online due to Covid-19
- 2020 Stochastics Seminar, Department of Mathematics, Georgia Institute of Technology (GaTech), online due to Covid-19
- 2020 Statistics Seminar, Marshall School of Business, University of Southern California (USC Marshall), online due to Covid-19
- 2020 Youth in High Dimensions: Machine Learning, High-dimensional statistics and Inference for the New Generation, *ICTP*, online due to Covid-19
- 2020 Discussant Talk, Mathematical Methods of Modern Statistics 2, CIRM, Luminy, online due to Covid-19
- 2020 Young Data Science Researchers Seminar, Swiss Federal Institute of Technology in Zurich (ETH Zurich), online due to Covid-19
- 2020 Mathematics of Data Science program, Society for Industrial and Applied Mathematics, online due to Covid-19
- 2019 Seminar, Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata
- 2019 Statistics Seminar, University of Minnesota
- 2019 Seminar in Econometrics, Harvard and MIT
- 2019 Department of Statistics and Data Science Seminar, Yale University
- 2019 Department of Statistics and Data Science Seminar, Carnegie Mellon University
- 2019 Statistics Seminar, University of Wisconsin-Madison
- 2019 Statistics Seminar, Columbia University
- 2019 Econometrics and Statistics Colloquium, The University of Chicago Booth School of Business
- 2019 Statistics Seminar, University of Chicago
- 2019 Statistics Seminar, Harvard University
- 2019 Operations Research and Financial Engineering Seminar, Princeton University
- 2019 Statistics Seminar, Rutgers University
- 2019 Statistics Seminar, The University of North Carolina at Chapel Hill
- 2019 Statistics Seminar, North Carolina State University
- 2019 Seminar, Decision Sciences Area, Duke Fuqua School of Business
- 2019 Statistics Seminar, Center for Data Science, New York University
- 2019 Statistics Seminar, The Wharton School at the University of Pennsylvania
- 2019 Statistics Seminar, University of Michigan

- 2019 Statistics Seminar, USC Marshall
- 2019 Frontiers in Computing and Mathematical Sciences, California Institute of Technology
- 2019 Statistics Seminar, University of Illinois at Urbana-Champaign
- 2019 Job Candidate Talk, Department of Mathematics, GaTech
- 2019 H.Milton Stewart School of Industrial and Systems Engineering Seminar, GaTech
- 2018 Statistics Seminar, University of California, Los Angeles (UCLA)
- 2018 Statistics Seminar, Cornell University
- 2018 **Department of Mathematics Colloquium**, University of California, San Diego
- 2018 Workshop in Biostatistics, Stanford University
- 2018 Workshop on Robust and High-Dimensional Statistics, Simons Institute for the Theory of Computing, University of California, Berkeley
- 2018 **Statistics Seminar**, *H. Milton Stewart School of Industrial and Systems* Engineering, GaTech
- 2018 Workshop on Algorithmic Challenges in Protecting Privacy for Biomedical Data, Institute for Pure and Applied Mathematics, UCLA
- 2014 Prasanta Chandra Mahalanobis Memorial Lecture, Indian Statistical Institute
- 2012 Debabrata Basu Memorial Lecture, Indian Statistical Institute

Funding

- 2024 The Eric and Wendy Schmidt Fund for Strategic Innovation, Project Title: "Unlocking Neurocognition and Beyond: Advancing Uncertainty Quantification for Interdisciplinary AI Systems"
- 2023-2024 Dean's Competitive Fund for Promising Scholarship, Project Title: "Assumption-Lean Transfer Learning in High Dimensions"
- 2021-2025 Division of Mathematical Sciences #2113426, National Science Foundation, Project Title: "Inference for Functionals in High-Dimensional Regression"
- 2020-2022 William F. Milton Fund, Harvard University, Project Title: "Prediction Methods for Remote Populations: Learning from Multiple Studies"
 - 2022 Radcliffe Institute for Advanced Study, Received funds for co-organizing workshop on "Exploring and Exploiting High-dimensional Phenomena in Statistical Learning and Inference"

Courses Taught

'22, '24, '25 Introduction to Supervised Learning

(S)

- '20 (F), '21 The Art of Communicating and Teaching Statistics, (co-instructed with (S) Xiao-li Meng)
- '21, '22, '23, Statistical Inference II
- '24, '25 (S)
- '21 (S & F), Timely Topics in Statistics, (co-instructed with Lucas Janson in Fall, 2021)
 '22 (F)

Student/Postdoc Advising and Collaboration

Postdoc

Samriddha Lahiry (2022-2024), Postdoctoral mentee, currently Visiting Fellow at the National University of Singapore

Manuel Saenz (2021-2023), *Joint work*, currently Assistant Professor of Statistics, University of Nottingham

Ph.D.

Anvit Garg (2028 expected), Thesis advisor

Kuanhao Jiang (2025 expected), Thesis advisor

Yufan Li (2025 expected), Thesis co-advisor

Yanke Song (2025 expected), Thesis advisor

Cathy Shyr (2022), *Joint work*, currently Postdoctoral Fellow at Vanderbilt University Dept. of Biostatistics

Undergraduate

Kenny Gu (2025 expected), Thesis advisor

Rachel Hong (2021), *Joint work*, currently Graduate Student at University of Washington Computer Science

Kevin Luo (2024), Thesis advisor (Thesis title: High-Dimensional Linear Interpolation for Structured Data, Winner of 2024 Harvard Hoopes Prize), currently Quantitative Trader at Hudson River Trading

Conlan Olson (2021), Thesis advisor (Thesis title: Algorithmic Fairness, Metric Embedding, and Metric Learning), currently Graduate Student at Columbia University Computer Science

Eddie Tu (2022), Thesis advisor (Thesis title: Invariance versus Adversarial Learning in Domain Generalization with Applications to Neuroscience), currently Quantitative Trader at Two Sigma

Professional Service Activities

Editorial Responsibilities

2024-present Associate Editor for Statistical Science

Reviewer for Grants

2022, 2024 Reviewer for National Science Foundation

2021 Reviewer for the William F. Milton Fund Competition, Harvard University

Reviewer for Journals/Conferences

2017-present Reviewer for Proceedings of the National Academy of Sciences, The Annals of Statistics, Journal of the Royal Statistical Society Series B, Journal of the American Statistical Association – Theory and Methods, Biometrika, Statistical Science, Statistica Sinica, Bayesian Analysis, IEEE Transactions on Information Theory, Operations Research, Journal of Machine Learning Research, Conference on Learning Theory (COLT), Conference on Neural Information Processing Systems (NeurIPS), Bernoulli, Information and Inference, Electronic Journal of Statistics, Journal of Statistical Planning and Inference, IEEE Journal on Selected Areas in Information Theory, Society of Industrial and Applied Mathematics (SIAM) Journal on Mathematics of Data Science

Service

- 2023-2025 Tenure-Track Committee, Harvard Statistics
- 2022-2023, Standing Committee on the S.M. Degree in Data Science, Harvard University 2024-2025
- 2024-2023
- 2020-2025 Co-organizer of Harvard Probabilitas Seminar Series, an interdisciplinary seminar series aimed at researchers at the interface of high-dimensional statistics, probability, and machine learning
- 2020-2022, Ph.D. Thesis Committee Member for Cathy Shyr, Zebin Wang, Harvard School
- 2022-2025 of Public Health
- 2020-2021, First Year Academic Advisor for Nicholas Barnfield, Nathan Cheng, Anvit Garg,
- 2022-2025 Tianle Liu
 - 2024 Program Committee Member for International Indian Statistical Association Conference, Cochin, India
- 2022-2024 President, IMS New Researchers Group
- 2022-2024 Ex-officio, IMS Outreach Committee and IMS Council
 - 2024 Organizer, Invited Session on "Recent Advances in Bayesian Inference", Bernoulli-IMS World Congress on Probability and Statistics (session on behalf of IMS New Researchers Group)
 - 2024 Organizer, Invited Session on "Modern Advances at the Interface of Statistical Learning and Inference", International Symposium on Nonparametric Statistics
 - 2024 External Examiner for Ph.D. Defense of Felix Kuchelmeister, ETH Zurich
 - 2024 Undergraduate Thesis Reader for Charumathi Badrinath and Kevin Luo
- 2022-2024 Oral Qualifying Committee member for Kuanhao Jiang, Biyonka Liang, Tyler Liu, Yanke Song, Longlin Wang, Elbert Du
- 2020-2023 Ph.D. Admissions Committee, Harvard University
 - 2023 Organizer, Invited Session on "Advances at the Interface of Statistics and Fair, Transferrable Machine Learning", IMS International Conference on Statistics and Data Science (session on behalf of IMS New Researchers Group)

- 2023 Organizer, Invited Session on "Modern Advances in Causal Inference", Joint Statistical Meetings (session on behalf of IMS New Researchers Group)
- 2023 Co-director, Workshop and Huddle on Learning and Inference from Structured Data: Universality, Correlations and Beyond, The Abdus Salam International Centre for Theoretical Physics (ICTP), Italy
- 2021-2022 Colloquium Chair/Co-Chair, Harvard Statistics
 - 2022 Co-Organizer of Workshop on "Exploring and Exploiting High-dimensional Phenomena in Statistical Learning and Inference", hosted by Radcliffe Institute for Advanced Study, Harvard University
 - 2022 Undergraduate Thesis Reader for Yash Nair and Eddie Tu
 - 2022 Dempster Prize Committee, Harvard Statistics
 - 2021 Undergraduate Thesis Reader for Conlan Olson
 - 2021 Organizer, IMS Invited Session on "Frontiers of High-Dimensional Statistics", Joint Statistical Meetings
 - 2021 Invited Speaker at Math Invitational for Girls 2021 Online Competition
 - 2021 Program Committee Member, Conference on Learning Theory
 - 2020 Mentor, "Rising Stars" Workshop on AI for Social Impact, Center for Research on Computation and Society (CRCS), Harvard SEAS
 - 2019 Panelist, Professional Development Panel on Interviewing, CRCS, Harvard SEAS