Harikrishna Narasimhan

Staff Research Scientist
Google LLC
1600 Amphitheatre Pkwy, Mountain View, CA 94043 Email: hnarasimhan@google.com

---

Research Interests

Efficient Inference for LLMs, Model Cascades, Loss Functions, Distillation, Learning to Rank, Constrained Optimization, Algorithmic Fairness

Employment

Staff Research Scientist 05/2025 - Present
Senior Research Scientist 11/2020 - 04/2025
Research Scientist 11/2018 - 10/2020
Google, Mountain View, USA

Post-doctoral Fellow 09/2015 - 08/2018
School of Engineering and Applied Sciences
Harvard University, Cambridge, USA
Advisor: Prof. David C. Parkes

Research Intern 06/2014 - 08/2014
Microsoft Research, Bangalore, India
Mentor: Dr. Prateek Jain

Education

Indian Institute of Science, Bangalore, India 08/2012 - 08/2015
Ph.D. in Computer Science (*Supported by a Google India PhD Fellowship*)
Advisor: Prof. Shivani Agarwal

Indian Institute of Science, Bangalore, India 08/2010 - 06/2012
M.E. Computer Science and Engineering

College of Engineering, Guindy, Chennai, India 07/2006 - 05/2010
B.E. Computer Science and Engineering

Journal Publications

Narasimhan, H., Ramaswamy, H.G., Tavker, S.K., Khurana, D., Netrapalli, P. and Agarwal, S. **Consistent Multiclass Algorithms for Complex Metrics and Constraints**. *Journal of Machine Learning Research (JMLR)*, 25(367):181, 2024.

Dutting, P., Feng, Z., Narasimhan, H., Parkes, D.C. and Ravindranath, S.S. **Optimal Auctions through Deep Learning: Advances in Differentiable Economics**. *Journal of the ACM (JACM)*, 71(1): 1-53, 2024.

Dutting, P., Feng, Z., Narasimhan, H., Parkes, D.C. and Ravindranath, S.S. **Optimal Auctions through Deep Learning**. Invited Research Highlight, *Communications of the ACM (CACM)*, 64(8):109-116, 2021.

Narasimhan, H. and Agarwal, S. **Support Vector Algorithms for Optimizing the Partial Area Under the ROC curve**. *Neural Computation*, 29(7):1919-1963, 2017.

Majumder, B., Baraneedharan, U., Thiyagarajan, S., Radhakrishnan, P., Narasimhan, H., Dhandapani, M., Brijwani, N., Pinto, D.D., Prasath, A., Shanthappa, B.U., Thayakumar, A., Surendran, R., Babu, G., Shenoy, A.M., Kuriakose, M.A., Bergthold, G., Horowitz, P., Loda, M., Beroukhim, R., Agarwal, S., Sengupta, S., Sundaram, M. and Majumder, P.K. **Predicting Clinical Response to Anticancer Drugs Using an Ex Vivo Platform that Captures Tumour Heterogeneity**. *Nature Communications* 6:6169, 2015.

Conference Publications

Lukasik, M., Chen, L., Narasimhan, H., Menon, A.K., Jitkrittum, W., Yu, F., Reddi, S.J., Fu, G., Bateni, M. and Kumar, S. **Bipartite Ranking From Multiple Labels: On Loss Versus Label Aggregation**. In the *42nd International Conference on Machine Learning (ICML)*, 2025. *To appear.*

Narasimhan, H., Jitkrittum, W., Rawat, A.S., Kim, S., Gupta, N., Menon, A.K. and Kumar, S. **Faster Cascades via Speculative Decoding**. In the *13th International Conference on Learning Representations (ICLR)*, 2025. [Outstanding Paper Honorable Mention Award] [Oral Presentation]

Lukasik, M., Meng, Z., Narasimhan, H., Menon, A.K., Chang, Y.-W., Yu, F. and Kumar, S. **Better Autoregressive Regression with LLMs**. In the *13th International Conference on Learning Representations (ICLR)*, 2025. [Spotlight presentation]

Lukasik, M., Narasimhan, H., Menon, A.K., Yu, F. and Kumar, S. **Regression-aware Inference with LLMs**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Narasimhan, H., Menon, A.K., Jitkrittum, W., Gupta, N., and Kumar, S. **Learning to Reject Meets Long-tail Learning**. In the *12th International Conference on Learning Representations (ICLR)*, 2024. [Spotlight presentation]

Gupta, N., Narasimhan, H., Jitkrittum, W., Rawat, A.S., Menon, A.K., and Kumar, S. **Language Model Cascades: Token-Level Uncertainty And Beyond**. In the *12th International Conference on Learning Representations (ICLR)*, 2024.

Narasimhan, H., Menon, A.K., Jitkrittum, W., and Kumar, S. **Post-hoc Estimators for Selective Classification and OOD Detection**. In the *12th International Conference on Learning Representations (ICLR)*, 2024.

Wei, J., Narasimhan, H., Amid, E., Chu, W.-S., Liu, Y., and Kumar, A. **Distributionally Robust Post-hoc Classifiers under Prior Shifts**. In the *11th International Conference on Learning Representations (ICLR)*, 2023.

Jitkrittum, W., Gupta, N., Menon, A.K., Narasimhan, H., Rawat, A.S., Kumar, S. **When Does Confidence-Based Cascade Deferral Suffice?**. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Wang, S., Narasimhan, N., Zhou, Y., Hooker, S., Lukasik, M., Menon, A.K. **Robust Distillation for Worst-class Performance**. In the *39th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

Narasimhan, H., Menon, A.K., Jitkrittum, W., Rawat, A.S., and Kumar, S. **Post-hoc estimators for learning to defer to an expert**. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Hiranandani, G., Mathur, J., Narasimhan, H., and Koyejo, O. **Quadratic Metric Elicitation with Application to Fairness**. 38th Conference on Uncertainty in Artificial Intelligence (UAI), 2022. [Oral presentation]

Jiang, H., Narasimhan, H., Bahri, D., Cotter, A. and Rostamizadeh, A. **Churn Reduction via Distillation**. In the *10th International Conference on Learning Representations (ICLR)*, 2022. [Spotlight Presentation]

Narasimhan, H. and Menon, A.K. **Training Over-parameterized Models with Non-decomposable Metrics**. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Hiranandani, G., Mathur, J., Narasimhan, H., Fard, M. M. amd Koyejo, O. **Optimizing Black-box Metrics with Iterative Example Weighting**. In the *38th International Conference on Machine Learning (ICML)*, 2021.

Kumar, A., Narasimhan, H., and Cotter, A. **Implicit Rate-constrained Optimization of Non-decomposable Objectives**. In the *38th International Conference on Machine Learning (ICML)*, 2021.

Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., and Jordan, M.I. **Robust Optimization for Fairness with Noisy Protected Groups**. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Narasimhan, H., Cotter, A., Zhou, Y., Wang, S., Guo, W. **Approximate Heavily-constrained Learning with Lagrange Multiplier Models**. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hiranandani, G., Narasimhan, H., and Koyejo, O. **Fair Performance Metric Elicitation**. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tavker, S.K., Ramaswamy, H.G., and Narasimhan, H. **Consistent Plug-in Classifiers for Complex Objectives and Constraints**. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jiang, Q., Adigun, O., Narasimhan, H., Fard, M.M., and Gupta M. **Optimizing Black-box Metrics with Adaptive Surrogates**. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Narasimhan, H., Cotter, A., Gupta, M., and Wang, S. **Pairwise Fairness for Ranking and Regression**. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

Narasimhan, H., Cotter, A., and Gupta, M. **Optimizing Generalized Rate Metrics with Three Players**. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019 [Oral Presentation]

Cotter, A., Narasimhan, H., and Gupta, M. **On Making Stochastic Classifiers Deterministic**. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. [Oral Presentation]

Zhao, S., Fard, M.M., Narasimhan, H. and Gupta, M. **Metric-optimized Example Weights**. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Dutting, P., Feng, Z., Narasimhan, H., Parkes, D.C. and Ravindranath, S.S. **Optimal Auctions through Deep Learning**. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. [Oral Presentation]

Narasimhan, H. **Learning with Complex Loss Functions and Constraints**. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Golowich, N., Narasimhan, H. and Parkes, D.C. **Deep Learning for Multi-Facility Location Mechanism Design**. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Feng, Z., Narasimhan, H. and Parkes, D.C. **Deep Learning for Revenue-Optimal Auctions with Budgets**. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2018.

Pan, W., Narasimhan, H., Kar, P., Protopapas, P. and Ramaswamy, H. **Optimizing the multiclass F-measure via biconcave programming**. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2016.

Li, S., Kar, P.K., Narasimhan, H., Chawla, S. and Sebastiani, F. **Stochastic optimization techniques for quantification performance measures**. In *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

Narasimhan, H., and Parkes, D.C. **A general statistical framework for designing strategy-proof assignment mechanisms**. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.

Narasimhan, H., Agarwal, S. and Parkes, D.C. **Automated mechanism design without money via machine learning**. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

Narasimhan, H., Parkes, D.C. and Singer, Y. **Learnability of influence in networks**. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Ahmed, S., Narasimhan, H. and Agarwal, S. **Bayes-optimal feature selection for supervised learning with general performance measures**. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

Narasimhan, H.*, Ramaswamy, H.G.*, Saha, A. and Agarwal, S. **Consistent multiclass algorithms for complex performance measures**. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.
(*both authors contributed equally to the paper)

Narasimhan, H., Kar, P. and Jain, P. **Optimizing non-decomposable performance measures: A tale of two classes**. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Kar, P., Narasimhan, H. and Jain, P. **Surrogate functions for maximizing precision at the top**. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Narasimhan, H.*, Vaish, R.* and Agarwal, S. **On the statistical consistency of plug-in classifiers for non-decomposable performance measures**. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.
(*both authors contributed equally to the paper)

Kar, P., Narasimhan, H., and Jain, P. **Online and stochastic gradient methods for non-decomposable loss functions**. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Agarwal, A., Narasimhan, H., Kalyanakrishnan, S., Agarwal, S. **GEV-canonical regression for accurate binary class probability estimation when one class is rare**. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

Saha, A., Dewangan, C., Narasimhan, H., Sriram, S., and Agarwal, S. **Learning score systems for patient mortality prediction in intensive-care units via orthogonal matching pursuit**. In *Proceedings of the 13th International Conference on Machine Learning and Applications (ICMLA)*, 2014.

Narasimhan, H. and Agarwal, S. **On the relationship between binary classification, bipartite ranking, and binary class probability estimation**. In *Advances in Neural Information Processing Systems (NIPS)*, 2013. [Spotlight Presentation]

Narasimhan, H. and Agarwal, S. **$\text{SVM}^{\text{tight}}_{\text{pAUC}}$: A new support vector method for optimizing partial AUC based on a tight convex upper bound**. In *Proceedings of the 19th ACM SIGKDD Conference on Knowledge, Discovery and Data Mining (KDD)*, 2013.

Menon, A. K., Narasimhan, H., Agarwal, S. and Chawla, S. **On the statistical consistency of algorithms for binary classification under class imbalance**. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

Narasimhan, H. and Agarwal, S. **A structural SVM based approach for optimizing partial AUC**. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

Book Chapter

Dutting, P., Feng, Z., Narasimhan, H., Parkes, D.C. and Ravindranath, S.S. **Machine Learning for Optimal Economic Design**. In *The Future of Economic Design*, Springer, 2019.

Preprints

Jitkrittum, W., Narasimhan, H., Rawat, A.S., Juneja, J., Wang, Z., Lee, C.-Y., Shenoy, P., Panigrahy, R., Menon, A.K and Kumar, S. **Universal Model Routing for Efficient LLM Inference.**. *Manuscript, 2025.* [arXiv:2502.08773]

Wang, C., Augenstein, S., Rush, K., Jitkrittum, W., Narasimhan, H., Rawat, A.S., Menon, A.K and Go, A. **Cascade-Aware Training of Language Models.**. *Manuscript, 2024.* [arXiv:2406.00060]

Cotter, A., Menon, A.K., Narasimhan, H., Rawat, A.S., Reddi, S.J. and Zhou, Y. **Distilling Double Descent**. *Manuscript, 2021.* [arXiv:2102.06849]

Professional Service

- **Senior Area Chair**: NeurIPS 2024, 2025
- **Area Chair:** ICML 2023, 2024, 2025; NeurIPS 2021, 2022, 2023
- **Conference Reviewing:** NeurIPS 2018, 2019, 2020; ICML 2020; ICLR 2021, 2022, 2023, 2024; FAccT 2021, 2022; COLT 2019; ACML 2021; STAC 2018; IJCAI 2016, IKDD CoDS 2016
- **Journal Reviewing:** Journal of Machine Learning Research, IEEE Transactions on Pattern Analysis and Machine Intelligence, ACM Transactions on Economics and Computation, Transactions on Machine Learning Research, Journal of Artificial Intelligence Research, Artificial Intelligence, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Cybernetics, Pattern Recognition Letters

Selected Awards and Achievements

- ICLR 2025 Outstanding Paper Honorable Mention Award for the paper Narasimhan et al., "*Faster Cascades via Speculative Decoding*".
- Reviewer awards:
  - Among top 10% of reviewers in NeurIPS 2020
  - Among top 33% of reviewers in ICML 2020
  - Among top 400 reviewers in NeurIPS 2019
  - Among top 200 reviewers in NeurIPS 2018.
- Google India PhD Fellowship in Machine Learning, 2013.
- Shell India Computational Talent Prize 2013 (SICTP) Gold Award.
- Computer Society of India (Bangalore Chapter) Medal for best M.E. student in computer science, Indian Institute of Science, 2012.

- Several awards during the bachelors program for academic excellence including one for Best Outgoing Student with Gold medal.

Patents

S. Kim, A.S. Rawat, W. Jitkrittum, H. Narasimhan, S. Reddi, N. Gupta, S. Bhojanapalli, A. Menon, M. Zaheer, T. Schuster, S. Kumar, T. Boyd, Z. Chen, E. Taropa, V. Kasivajhula, T. Strohman, M. Baeuml, L. Schelin, and Y. Huang. **Dynamic selection from among multiple candidate generative models with differing computational efficiencies**, US Patent, US20240311405A1, 2024.

H. Narasimhan, W. Jitkrittum, A.K. Menon, A.S. Rawat, and S. Kumar. **Performing classification tasks using post-hoc estimators for expert deferral**, US Patent, US20240135254A1, 2024.

A. Kumar, H. Narasimhan, and A.S. Cotter, **Systems and methods for implicit rate-constrained optimization of non-decomposable objectives**, US Patent, US20220398506A1, 2022.

Selected Talks And Outreach

- Invited talk on "Learning to Abstain and Beyond" at the 1st AAAI Workshop on Deployable AI (DAI), 2023.
- Co-organized tutorial on "Deep AUC Maximization: From Algorithms to Practice" at CVPR 2022.
- Guest lecture on"Fairness Goals through Constrained Optimization" in the AI and Ethics course at IIT Kharagpur, 2020.
- Co-authored Google blogpost on "Setting Fairness Goals with the TensorFlow Constrained Optimization Library", 2020.
- Talk on "Simple Algorithms for Complex Learning Problems" at Google, Mountain View, 2018.

References

Available upon request.