Parameswaran Raman

640 Epic Way, San Jose, CA 95134

☎ (408) 306 4462 ✉ parameshr@gmail.com https://paramsraman.github.io/

Research Interests

Large Scale Machine Learning, Optimization, Distributed Learning, Deep Neural Networks, Ranking & Recommender Systems, NLP.

Education

PhD - Computer Science, UC Santa Cruz **Aug 2013 - Dec 2019**
(Transferred from Purdue University 2013 - 2014)
Advisor: Prof. S.V.N. Vishwanathan
Thesis Title: *Hybrid-Parallel Parameter Estimation for Frequentist and Bayesian Models*
Thesis Committee: S.V.N. Vishwanathan, Manfred K. Warmuth, David P. Helmbold

Masters - Computer Science, Georgia Institute of Technology **Aug 2009 - May 2011**

MSc (Integrated) - Software Engineering, PSG College of Technology, India **2003 - 2008**

Research & Professional Experience

Applied Scientist - *Amazon AWS AI* **Apr 2021 - Present**

- Research and development of large-batch training methods (optimization algorithms as well as distributed infrastructure) to enable higher throughput and downstream performance using large batch sizes on transformer models.
- Research and development of 3D parallelism approaches (data, pipeline and tensor model parallel) to pre-train large-scale transformer models on Amazon data involving billions of parameters. Work also involves benchmarking and debugging popular open-source deep-learning libraries such as DeepSpeed, Megatron and Fairscale and providing occasional fixes to the open-source community.
- Mentored interns on research projects
  - Pipeline parallelism for large-sequence transformer models
  - Developing a memory efficient error-feedback method using gradient compression
  - Developing custom model parallelism approaches for large scale multi-model transformer models

Applied Scientist - *Amazon Alexa AI* **Apr 2020 - Apr 2021**

- Designed a unified DNN model architecture for handling span-based and picklist categorical attributes in Alexa Conversations dialog models.
- Developed an offline and online (sagemaker) pipeline for end-end evaluation of dialog models.
- Helped the team onboard a new cluster access and management system. Evaluated data transfer, training and evaluation use-cases and prepared a detailed onboarding document.
- Worked on migrating modeling and evaluation features from older PyTorch code base to newer Deep Learning based NLU framework (based on MXNet) for Alexa Conversations platform.
- Mentored an intern on Pre-training Graph Neural Networks on natural language data for downstream NLP tasks.

Graduate Student Researcher - *UC Santa Cruz* **Aug 2014 - March 2020**

- Developed a scalable distributed algorithm (DS-FACTO) to scale Factorization Machines to large data and large number of features.
- Developed a new distributed asynchronous bayesian inference algorithm for large scale mixture models (ESVI: Extreme Stochastic Variational Inference), which is both model and data parallel.

- Developed a scalable, distributed stochastic optimization algorithm for multinomial logistic regression (DS-MLR) on massive datasets with large number of examples and classes. DS-MLR is hybrid-parallel (de-centralizes both data and the model simultaneously), lock-free and asynchronous.

Graduate Student Researcher - *Purdue University* **Aug 2013 - May 2014**

- Developed a new learning to rank algorithm (RoBiRank) inspired by Robust Binary Classification, which directly bounds NDCG. Extended it to the Latent Collaborative Retrieval setting and developed a distributed stochastic optimization algorithm that can scale to large datasets. RoBiRank was deployed and tested at LinkedIn (internship) on part of the live traffic. Obtained 6.4% lift in CTR@P1 and 2.5% reduction in abandonment of user sessions.

Applied Scientist Intern - *Amazon AWS AI* **Summer 2017**

- Worked with Recommendations and Search Team at Twitch to develop ranking models for video clip recommendation. Investigated and developed a proof-of-concept implementation for a temporal deep recommender system (Hybrid using Matrix Factorization and LSTMs) to model the sequence of temporal effects in users and items.

Research Intern - *Adobe Research (Systems Technology Lab)* **Summer 2016**

- Developed models to cluster user-behavior in Adobe analytics data using both click (user url) as well as content (user meta-data) information.

Research Intern - *Microsoft (Cloud and Information Services Lab)* **Summer 2015**

- Researched the problem of extrapolating learning curves in machine learning - to study if it was feasible to use information from the models learnt on various sizes of small bites of data to extrapolate performance of the algorithm on the full data. Implemented the prototype.

Research Intern - *LinkedIn (Search Relevance - SNA)* **Summer 2014**

- Explored learning based approaches to resolve sample bias and position bias present in learning to rank systems. Developed a new framework to combine ranking models incrementally.

Graduate Student Researcher - *Georgia Tech (Sonification Lab)* **Aug 2009 - May 2011**

Software Engineer - *Yahoo!, Sunnyvale* **Jul 2011 - Jul 2013**

- Worked for the Personalization group on an entity detection platform used by all personalization services. Built a Knowledge Graph from scratch to power Yahoo! search products.
- Worked on the Web of Objects project, to create a semantic knowledge base of entities to enable personalization. Designed features for Entity Matching models and wrote tools to evaluate them.
- Worked on Apache Oozie (job scheduler for Hadoop), implementing key features and fixing bugs.

Software Engineering Intern - *Intel, Chandler* **Summer 2010**

Software Engineer - *ThoughtWorks, Bangalore* **Jun 2008 - Jul 2009**

Publications

- Bingcong Li, Shuai Zheng, Parameswaran Raman, Anshumali Shrivastava, Georgios B Giannakis. **"Contractive error feedback for gradient compression"**, *Preprint on Openreview 2021*.
- Parameswaran Raman, S.V.N. Vishwanathan. **"DS-FACTO: Doubly-Separable Factorization Machines"**, *Tech Report, arXiv 2020*.
- Parameswaran Raman, Sriram Srinivasan, Shin Matsushima, Xinhua Zhang, Hyokun Yun, S.V.N. Vishwanathan. "**Scaling Multinomial Logistic Regression via Hybrid-Parallelism**," *KDD* 2019. **Accepted as Oral Presentation (9.16 % acceptance rate)**.

- Parameswaran Raman*, Jiong Zhang*, Shihao Ji, Hsiang-Fu Yu, S.V.N. Vishwanathan, Inderjit S. Dhillon. "**Extreme Stochastic Variational Inference: Distributed and Asynchronous** ," *AISTATS* 2019.

- Hyokun Yun, Parameswaran Raman, S.V.N. Vishwanathan. "**Ranking via Robust Binary Classification and Parallel Parameter Estimation in Large-Scale Data**," *NIPS*. 2014.

- Mariheida Córdova Sánchez, Parameswaran Raman, Luo Si, Jason Fish. "**Relevancy Prediction of Micro-blog Questions in an Educational Setting**," in *Proceedings of the 7th International Conference on Educational Data Mining, EDM*. 2014.

- Parameswaran Raman, Jiasen Yang. "**Optimization on the Surface of the (Hyper)-Sphere**" Tech Report *arXiv: 1909.06463*. 2014.

- Myounghoon "Philart" Jeon, Benjamin Davison, Jeff Wilson, Parameswaran Raman, Bruce N. Walker. "**Advanced Auditory Menus for Universal Access to Electronic Devices**," in *Proceedings of CSUN International Technology & Persons with Disabilities Conference*. 2010.

- Parameswaran Raman, Benjamin Davison, Myounghoon "Philart" Jeon, Bruce N. Walker. "**Reducing repetitive development tasks in auditory menu displays with the auditory menu library**," in *Proceedings of the 16th International Conference on Auditory Display (ICAD)*. 2010.

- Parameswaran Raman, Narayanan Ramakrishnan, Manohar Ganesan, Gourab Kar, Dr Gregory D. Abowd. "**PiX-C: Express and Communicate (Augmenting Communication with Visual Input for Children in the Autism Spectrum)**," in *Poster presented at ACM Symposium on User Interface Software and Technology (UIST)*. 2010.

- Mary Magdalene Jane, Parameswaran Raman, Maytham Safar, Nadarajan R. "**PINE-guided cache replacement policy for location-dependent data in mobile environment**," in *Proceedings of the First international conference on Pervasive Technologies Related to Assistive Environments, PETRA*. 2008.

- Parameswaran Raman, Raghavendra Prasad, Nadarajan R, Mary Magdalene Jane. "**Weighted Angular Distance Based Cache Replacement Strategy for Location-Dependent Data in Wireless Environment**," in *Proceedings of the DCCA Conference, Jordan*. 2007.

Academic Services

- Reviewing activities
  - Conferences: UAI 2014, AISTATS 2015, COLT 2015, AISTATS 2016, ICML 2016, NeurIPS 2018, ICML 2019, NeurIPS 2019, ICML 2020, ICLR 2021, ICLR 2022
  - PC member: AAAI 2020, 2021
  - Journals: JMLR 2015, TPAMI 2015
- Book Chapter Reviewer: *"Mathematics for Machine Learning", Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong*, Cambridge University Press, 2020, ISBN: 9781108455145
- Instructor for *Tools of the Trade Bootcamp Series* at UC Santa Cruz (Topics: git, LaTeX, Editors, Make, Unix, Shell, etc.) in Fall 2015.
- Teaching Assistant for *CMPS 242 - Grad Level Machine Learning* course at UC Santa Cruz (Instructor: Prof. S.V.N. Vishwanathan) in Fall 2016.

Talks

- IBM Research Almaden (Jan 2020)
  Hybrid-Parallel Parameter Estimation for Machine Learning

- Fiddler.ai (Sep 2019)
  Scaling Multinomial Logistic Regression via Hybrid Parallelism
- Oral Presentation (KDD 2019)
  Scaling Multinomial Logistic Regression via Hybrid Parallelism
- PSG College of Technology, Coimbatore India (Summer 2018)
  Recipes for PhD - A Machine Learning Perspective
- AMS 250 - High Performance Computing course at UC Santa Cruz (Spring 2018)
  Distributed Machine Learning: Approaches and Challenges
- Machine Learning Lab Seminar, UC Santa Cruz
  - *Extreme Stochastic Variational Inference (ESVI): Distributed Inference for Large Scale Mixture Models*
  - *Large-Scale Distributed Bayesian Matrix Factorization using Stochastic Gradient MCMC*
  - *Cover Trees for Nearest Neighbor Search*
  - *Tutorial on Variational Inference*

Honors & Awards

- Travel award for Neural Information Processing Systems (NIPS), 2014
- Travel award for Tripods Summer School on Foundations of Data Analysis at UW Madison, 2018
- Travel award for Machine Learning in Science and Engineering (MLSE) conference at CMU, 2018
- Winner of Facebook Hackathon at Georgia Tech & finalist at FB HQ, 2010
- Finalist at UIST Student Innovation Contest, 2010

Open Source Software

- DS-MLR (Hybrid-Parallel Multinomial Logistic Regression)
- ESVI (Hybrid-Parallel Variational Inference for Mixture Models)
- RoBiRank (Robust and Scalable Ranking Algorithm for Large Data)

Skills

- *Programming:* C/C++, Python, Java, Matlab
- *Deep Learning:* Tensorflow, MXNet
- *Parallel Computing:* MPI, Open MP, Intel TBB
- *Data Processing:* Hadoop, Pig, Apache Spark

References

Available upon request.