

Jerry Chee

Department of Computer Science
Cornell University

JerryChee@cs.cornell.edu
Jerry-Chee.github.io

I am interested in developing machine learning methods to meet the needs of practitioners.

Education	Cornell University	Ithaca, NY
	Ph.D. in Computer Science Advisor: Chris De Sa	2019 - 2025 (expected)
	University of Chicago	Chicago, IL
	B.S. in Computational and Applied Mathematics Advisor: Panos Toulis	2013 - 2017
Publications	J. Chee , Y. Cai, V. Kuleshov, C. De Sa. <i>QuIP: 2-Bit Quantization of Large Language Models with Theoretical Guarantees</i> In <i>NeurIPS 2023 (Spotlight)</i>	
	J. Chee , H. Kim, P. Toulis. “Plus/minus the learning rate”: Easy and Scalable Statistical Inference with SGD. In <i>AI and Statistics 2023</i>	
	J. Chee , M. Renz, A. Damle, C. De Sa. <i>Model Preserving Compression for Neural Networks</i> . In <i>NeurIPS 2022</i>	
	J. Chee , S. Braun, V. Gopal, R. Cutler. <i>Performance Optimizations on U-Net Speech Enhancement Models</i> . In <i>IEEE Multimedia Signal Processing 2022</i>	
	C. Yang, Z. Wu, J. Chee , C. De Sa, M. Udell. <i>How Low Can We Go: Trading Memory for Error in Low-Precision Training</i> . In <i>ICLR 2022</i>	
	J. Chee , P. Li. <i>Understanding and Detecting Convergence for Stochastic Gradient Descent</i> . In <i>IEEE Big Data 2020</i>	
	J. Chee , P. Toulis. <i>Convergence Diagnostics for Stochastic Gradient Descent</i> . In <i>AI and Statistics 2018 (Oral)</i>	
Talks	Statistical Properties of Stochastic Gradient Descent	Denver, CO
	<i>Joint Statistics Meeting</i> , with Panos Toulis.	Jul 2019
	Convergence Diagnostics for Stochastic Gradient Descent	Canary I.
	<i>AISTATS 2018</i> , with Panos Toulis.	Apr 2018
Projects	QuIP#: Incoherence Processing with Lattice Codebooks	
	- with A. Tseng, C. De Sa, V. Kuleshov	
	Better quantization with codebooks, faster inference with custom CUDA kernels.	
	Harm-Mitigation in Recommender Systems (In submission)	
	- with S. Ernala, S. Kalayanaraman, S. Ioannidis, S. Dean, U. Weinsberg	
	Study recommender policies which mitigate user engagement with harmful content.	
	Predicting lincRNA functionality in short and long ORF (In preparation)	
	- with C. Railey, C. De Sa, A. Nelson	
	Predict protein coding ability of lincRNA using deep recurrent NN models.	

Industry Experience	Meta , Core Data Science <i>Research Engineer Intern</i>	Menlo Park, CA Jun–Sept 2022
	<ul style="list-style-type: none"> • Prototyped deep learning-based metric to estimate the likelihood a user would interact with borderline harmful content based on previous interaction history. • Compiled requisite datasets using SQL, performed data analysis and visualization in notebooks, and trained distributed DNNs at scale. 	
	Amazon , Supply Chain Optimization Technologies <i>Applied Scientist Intern</i>	Seattle, WA Dec 2021–May 2022
	<ul style="list-style-type: none"> • Estimated $12\times$ training speedup for a causal inference model used to estimate the value of in-stock items on Amazon.com. • Saved and reused repeated computation via repeated linear regressions with common set of controls. 	
	Microsoft , IC3-AI <i>Intern</i>	Redmond, WA Jun–Sept 2021
	<ul style="list-style-type: none"> • $7\times$ inference speedup of deep background noise suppression models used real-time in Teams. • Identified and implemented model compression methods supported by the neural network inference engines ONNX Runtime, CoreML, and TFLite. 	
	Baidu , Cognitive Computing Lab <i>Research Intern</i>	Bellevue, WA Mar–Jul 2019
	<ul style="list-style-type: none"> • Developed statistical convergence tests for variants of stochastic gradient descent with momentum and gradient compression. • Utilized multi-task learning to increase the available training data in order to improve the predictive performance of graph neural networks. 	
	McKinsey & Company <i>Senior Analytics Fellow</i>	Boston, MA Oct 2017 - Feb 2019
	<ul style="list-style-type: none"> • Implemented data science solutions at client organizations, working closely with business leaders and domain experts. • Led several data science initiatives in predictive maintenance for the network technology division of a top telecommunications company. <ul style="list-style-type: none"> – Utilized a cost (of true positive, false positive, etc.) analysis for selecting the prediction target and implementation strategy which maximized business impact and modeling feasibility. – Built classification models for network and customer service use cases. 	
Teaching	TA, CS 4780/5780: Machine Learning for Intelligent Systems	Fall 2019
	TA, CS 4787: Principles of Large-Scale Machine Learning	Spring 2020
	TA, CS 6787: Advanced Machine Learning Systems	Fall 2020
Outreach	Skype A Scientist Volunteer Video call with classrooms across the country to help educate students about research in computer science and career options as a quantitative scientist.	Apr 2020-May 2021
Other Information	Programming: Python (PyTorch), SQL, R (Rcpp), C (MPI)	
	Languages: Chinese (Limited oral proficiency)	