Ekansh Sharma

Contact Information

Email: ekanshs@cs.toronto.edu | Website: ekansharma.org | GitHub: github.com/ekanshs
Google Scholar: scholar.google.com Mountain View, CA, USA

Research Profile

Machine learning researcher and engineer (Ph.D., University of Toronto) specializing in large-scale sequential modeling, optimization geometry, low-precision training numerics, and statistical inference under noisy, non-stationary feedback.

Professional & Research Experience

Google, Mountain View, CA, USA *September 2025 – Present*
Machine Learning Engineer — YouTube Shorts

- **Generative Sequential User Modeling (*TokenMinds*, arXiv:2606.25147):** Co-authored and deployed an LLM-adapted encoder-decoder user model (370M MoE encoder / 370M dense decoder in JAX) unifying short-form, long-form, and search histories via hierarchical RQ-VAE Semantic IDs; serves billions of users asynchronously (1.44M read QPS, 96.4% cache hit rate), cutting upstream training and serving compute by 50% and 31% while driving statistically significant live A/B gains on YouTube Shorts (+0.13% Satisfied Engagement, +0.11% Engaged Users, +0.33% Freshness Engagement).
- **Long-Context Compression & Representation Stability:** Extended sequential user modeling to long interaction horizons using soft-token compression, sparse cross-attention bottlenecks, and multi-resolution recency compression. Established longitudinal representation-stability diagnostics (cosine drift, perturbation invariance, split-history self-retrieval, and effective-rank monitoring) and non-autoregressive auxiliary reconstruction objectives to prevent representation collapse under continuous daily training.
- **Causal Point-Process Objectives, Variance Reduction & Cold-Start Estimation:** Designed multi-task sequence objectives for noisy implicit-feedback streams, incorporating right-censored survival/hazard modeling, count-process supervision, deterministic importance/loss weighting, and C++ streaming input filters with strict request-time causality checks against look-ahead label leakage; developed empirical-Bayes shrinkage representations for cold-start candidate scoring.
- **Optimization Symmetries & Low-Precision Training Numerics:** Investigated orthogonal gauge symmetries and outlier-mitigation techniques for low-precision (`FP8/INT4`) quantization-aware training under matrix-sign and adaptive optimizers; profiled end-to-end XLA/TPU HBM, VMEM, and operator-fusion rooflines on large-scale multi-task ranking architectures.
- **Scaling Laws & Experimentation Infrastructure:** Conducted Iso-FLOPS capacity, batch-size scaling, Maximal Update Parameterization (μP), and low-compute proxy transfer studies to predict full-scale ranking performance; led multi-launch model integration cycles and built automated training-lifecycle governance and paired statistical evaluation tooling for continuous TPU training pipelines.

Vector Institute & University of Toronto, Toronto, ON, Canada *March 2018 – August 2025*
Graduate Researcher (Adviser: Prof. Daniel M. Roy)

- **Loss-Landscape Geometry & Non-Local Model Merging:** Established simultaneous linear mode connectivity modulo permutation symmetries across neural network populations (ECML-PKDD 2024; CPAL Spotlight 2024); identified deep-layer variance collapse in non-local multi-task model merging across distinct foundation models and open-sourced task-specific statistical repair (*TACT*, arXiv:2410.12766, code).
- **Sparse Optimization Dynamics:** Analyzed normalization-induced gradient skew in sparse training from scratch (*SparseOpt*, ICML 2026) and weight-symmetry permutation alignment for random-initialization lottery-ticket masks (ICML 2025).
- **Nonasymptotic Monte Carlo & Heavy-Tailed Random Graphs:** Derived compositional nonasymptotic error bounds for nested MCMC in probabilistic programs (arXiv:1912.06791) and consistent bootstrap tail-index estimators for sparse exchangeable power-law graphs (*Electronic Journal of Statistics*, 2021).

Amazon.com (Sponsored Display Ads), Edinburgh, UK *January 2021 – May 2021*
Applied Scientist Intern

- Developed counterfactual off-policy evaluation estimators for ad-selection policies to quantify conversion lift while reducing live A/B exploration cost; evaluated Upper Confidence Bound (UCB) contextual bandit algorithms in production A/B experiments measuring click-conversion rate.

Education

University of Toronto, Toronto, ON, Canada
Ph.D. in Computer Science, conferred November 2025 (Adviser: Prof. Daniel M. Roy)
Thesis: *Exploiting linear connectivity modulo permutations in neural networks for parsimonious machine learning*

M.Sc. in Computer Science, conferred 2018 (Adviser: Prof. Daniel M. Roy; Thesis: *Sparse exchangeable bipartite graphs*)

B.A.Sc. with Honors in Electrical Engineering (Minor in Robotics and Mechatronics), conferred 2016

Selected Publications & Preprints

Liu Q., Yan B., Liu Y., Roh Y., **Sharma E.**, Yin L., Olowo E., Li Y., Jiang Y., Uribe D., Aggarwal S., Wu S., Hao Y., Kedigehalli V., Tsai M.-H., Heldt L., Hong L., Wei L., Yi X. "TokenMinds: Pretrained User Tokens and Embeddings for User Understanding in Large Recommender Systems." *arXiv:2606.25147* (2026).

Adnan M., Jain R., Jacobs T., **Sharma E.**, Krishnan R. G., Burkholz R., Ioannou Y. "SparseOpt: Addressing Normalization-induced Gradient Skew in Sparse Training." *International Conference on Machine Learning (ICML)*, 2026 (*arXiv:2605.27541*).

Adnan M., Jain R., **Sharma E.**, Krishnan R. G., Ioannou Y. "Sparse Training from Random Initialization: Aligning Lottery Ticket Masks using Weight Symmetry." *International Conference on Machine Learning (ICML)*, 2025 (*arXiv:2505.05143*).

Sharma E., Roy D. M., Dziugaite G. K. "The Non-Local Model Merging Problem: Permutation Symmetries and Variance Collapse." *arXiv:2410.12766* (2024).

Sharma E., Kwok D., Denton T., Roy D. M., Rolnick D., Dziugaite G. K. "Simultaneous Linear Connectivity of Neural Networks Modulo Permutation." *ECML-PKDD*, 2024; *CPAL Spotlight Track*, 2024 (*arXiv:2404.06498*).

Naulet Z., Roy D. M., **Sharma E.**, Veitch V. "Bootstrap estimators for the tail-index and for the count statistics of graphex processes." *Electronic Journal of Statistics*, 15(1), 282–325, 2021.

Sharma E., Roy D. M. "Approximations in Probabilistic Programs: A Compositional Nonasymptotic Analysis of Nested MCMC." *ProbProg*, 2020 (*arXiv:1912.06791*).

Veitch V., **Sharma E.**, Naulet Z., Roy D. M. "Exchangeable modelling of relational data: checking sparsity, train-test splitting, and sparse exchangeable Poisson matrix factorization." *arXiv:1712.02311* (2017).

Technical Skills

Languages & Systems: Python, C++, C, SQL, Bash; multi-slice TPU/GPU distributed training, custom C++ streaming data & sparse coordinate kernels, XLA/Pallas profiling, HBM/VMEM roofline analysis.

ML & Optimization: JAX/Flax, PyTorch, TensorFlow; LLM/Transformer sequence modeling, non-Euclidean/matrix-sign & sparse optimization, low-precision (`FP8`/`INT4`) training numerics, model merging, μP scaling.

Statistical & Quantitative Methods: Survival & point-process hazard modeling, counterfactual and off-policy evaluation, contextual bandits, empirical-Bayes shrinkage, heavy-tailed asymptotics, nonasymptotic concentration, MCMC, low-SNR paired/bootstrap inference.

Honors & Service

Awards: Vector Research Grant (2018–2025); Dept. of CS 50th Anniversary Graduate Scholarship (2017); U of T Excellence Award (2013, 2015); Dean's Honors List (2011–2016).

Academic Service: Reviewer for NeurIPS (2021–2023, 2025–2026), ICML (2023–2025), and ICLR (2020–2023).