Youngsuk Park

Email: youngsuk@cs.stanford.edu
Phone: +1 (650) 422-8541
Homepage: http://cs.stanford.edu/~youngsuk/
LinkedIn: https://www.linkedin.com/in/y-park/

Research Interests

Hardware-Aware Model Architectures: fine grained scaling laws and architecture design for high-throughput
Efficient Training: Low-precision training, system-aware optimizers, and distributed training
Post-training and Alignment: RLVR-based post-training, robustness reward modeling
Generative AI for Systems: DSL kernel synthesis via inference-time scaling and RLVR

Education

Ph.D. in Electrical Engineering, Stanford University, 2020
Advisors: Stephen P. Boyd and Jure Leskovec
Dissertation: *Topics in Convex Optimization for Machine Learning*

M.S. in Electrical Engineering, Stanford University, 2016

B.S. in Electrical Engineering (Minor in Mathematics), KAIST, 2013
Summa Cum Laude

Professional Experience

Senior Applied Scientist & Research Lead, Amazon Web Services AI, Mar. 2023–Present

- Lead Core Algorithm team advancing scalable LLM training and inference for AWS Trainium
- Manage research organization of 14+ applied scientists, Amazon scholars, and research interns
- Pioneer innovations in quantization, structured sparsity, and hardware-aware modeling
- Deploy foundation models across Amazon Bedrock, AGI, and Anthropic partnerships
- Research on DSL kernel synthesis (GPU Triton, Trainium NKI) using RLVR and inference-time scaling

Applied Scientist II, Amazon Web Services AI Labs, Jun. 2020–Mar. 2023

- Technical lead for time series forecasting and foundation model development
- Led third-party validation of foundation models on AWS Trainium accelerators

Research Intern, Adobe Research, Jun.–Sept. 2019
Focus: Reinforcement learning for continuous control and cloud resource management

Research Intern, Criteo AI Lab, Jun.–Sept. 2018
Focus: Off-policy reinforcement learning for recommendation systems

Research Intern, Bosch Center for AI, Jun.–Sept. 2017

Publications

Preprint/Under Review (3)

[1] LLM RL SYS J. Woo, S. Zhu, A. Nie, Y. Wang, **Y. Park**†. **TritonRL: Training LLMs to Think and Code Triton Without Cheating.** Under review.

[2] LLM SYS J. Yoo, R. Saha, T. Yu, **Y. Park**†. **Modular Kernel Evolution: LLM-driven Kernel Synthesis for Domain Specific Language.** Under review.

[3] LLM SYS **Y. Park**†, K. Budhathoki, L. Chen, J. Kübler, J. Huang, M. Kleindessner, Y. Wang, G. Karypis. **Accelerate LLM Inference via 4:8 Semi-structured Sparsity on AWS Trainium.** Under review.

Peer-reviewed Publications (36)

[4] OPT LLM A. Khaled, K. Ozkara, T. Yu, **Y. Park**†. **MuonBP: Faster Muon Optimizer via Block-Periodic Orthogonalization.** *International Conference on Learning Representations (ICLR)*, 2026.

[5] LLM SYS S. Bian, T. Yu, **Y. Park**†. **Scaling Laws Meet Model Architecture: Toward Inference-Efficient LLMs.** *International Conference on Learning Representations (ICLR)*, 2026.

[6] LLM RL H. Liu, J. Huang, **Y. Park**, Y. Wang. **Not-a-Bandit: Provably No-Regret Drafter Selection in Speculative Decoding for LLMs.** *International Conference on Learning Representations (ICLR)*, 2026.

[7] ML LLM J. Kim, R. Saha, M. Sung, **Y. Park**. **Demystifying Transition Matching: When and Why It Can Beat Flow Matching.** *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026.

[8] LLM OPT Q. Hong, C. Chung, **Y. Park**, M. Hong. **RoSTE: An Efficient Quantization-Aware Supervised Fine-Tuning Approach for Large Language Models.** *International Conference on Machine Learning (ICML)*, 2025.

[9] LLM H. Liu, R. Saha, **Y. Park**, Y. Wang. **ProxSparse: Regularized Learning of Semi-Structured Sparsity Masks for Pretrained LLMs.** *International Conference on Machine Learning (ICML)*, 2025.

[10] TS LLM L. Masserano, A. Ansari, C. Faloutsos, M. Mahoney, A. Wilson, **Y. Park**, Y. Wang. **Enhancing Foundation Models for Time Series Forecasting via Wavelet-based Tokenization.** *International Conference on Machine Learning (ICML)*, 2025.

[11] LLM SYS A. Tseng, T. Yu, **Y. Park**†. **Training LLMs with MXFP4.** *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

[12] LLM OPT K. Ozkara, T. Yu, **Y. Park**†. **Stochastic Rounding for LLM Training: Theory and Practice.** *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

[13] ML B. Kevton, B. Oreshkin, **Y. Park**, R. Song. **Contextual Posterior Sampling with a Diffusion Model Prior.** *Neural Information Processing Systems (NeurIPS)*, 2024.

[14] LLM SYS **Y. Park**†, K. Budhathoki, L. Chen, J. Kübler, J. Huang, M. Kleindessner, J. Huan, V. Cevher, Y. Wang, G. Karypis. **Survey: Inference Optimization of Foundation Models on AI Accelerators.** *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

[15] LLM OPT T. Gautam, **Y. Park**†, H. Zhou, P. Ramen. **Variance-reduced Zeroth-Order Methods for Fine-Tuning LLMs with just Forward Pass.** *International Conference on Machine Learning (ICML)*, 2024.

[16] LLM OPT T. Yu, G. Gupta, K. Gopalswamy, **Y. Park**, J. Huan, R. Diamond. **Collage: Light-Weight Low-Precision Strategy for LLM Training.** *International Conference on Machine Learning (ICML)*, 2024.

[17] LLM UQ C. Marx, W. Ha, C. Bock, J. Huan, **Y. Park**†. **Eliciting Calibrated Uncertainties from Language Models.** *Amazon Machine Learning Conference (AMLC)*, 2024.

[18] ML H. Ding, Y. Ma, **Y. Park**, A. Deoras, H. Wang, B. Kveton. **Trending Now: Modeling Trend Recommendations.** *ACM International Conference on Recommender Systems (RecSys)*, 2023.

[19] ML TS H. Hasson, D. Maddix, Y. Wang, **Y. Park**. **Theoretical Guarantees of Learning Ensembling Strategies with Applications to Time Series Forecasting.** *International Conference on Machine Learning (ICML)*, 2023.

[20] ML TS L. Lui, **Y. Park**†, N. Hoang, H. Hasson, J. Huan. **Robust Multivariate Time-Series Forecasting: Adversarial Attacks and Defense Mechanisms.** *International Conference on Learning Representations (ICLR)*, 2023.

[21] UQ ML C. Marx, **Y. Park**†, H. Hasson, Y. Wang, J. Huan, S. Ermon. **But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI.** *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[22] TS RL Y. Ding, **Y. Park**†, K. Gopalswamy, Y. Wang, J. Huan. **Dynamic Ensembling for Probabilistic Time Series Forecasting: Reinforcement Learning Approach.** *KDD Time Series MILETS Workshop*,

2023.

[23] ML TS X. Jin, **Y. Park†**, D. Maddix, Y. Wang. **Domain Adaptation for Time Series Forecasting via Attention Sharing.** *International Conference on Machine Learning (ICML)*, 2022.

[24] TS UQ **Y. Park†**, D. Maddix, J. Gasthaus, Y. Wang. **Learning Quantile Functions without Quantile Crossing for Distribution-free Time Series Forecasting.** *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[25] ML TS UQ T. Yoon, **Y. Park†**, E. Ryu, Y. Wang. **Robust Probabilistic Forecasting via Randomized Smoothing.** *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022.

[26] TS UQ K. Kan, F. Aubet, T. Januschowski, **Y. Park**, K. Bendis, J. Gasthaus. **Multivariate Quantile Functions for Forecasting.** *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022. **Selected as oral, 2.6% of all submissions.**

[27] TS ML L. Masserano, S. Rangapuram, R. Nirwan, S. Kapoor, **Y. Park**, M. Bohlke-Schneider. **Adaptive Sampling for Probabilistic Forecasting Under Distribution Shift.** *NeurIPS Workshop on Distribution Shift*, 2022.

[28] TS LLM X. Zhang, X. Jin, K. Gopalswamy, **Y. Park**, D. Maddix, Y. Wang. **First De-Trend then Attend: Rethinking Attention for Time-Series Forecasting.** *NeurIPS Workshop on Attention Models*, 2022.

[29] TS J. Zhang, **Y. Park**, H. Hasson, D. Maddix, D. Roth, Y. Wang. **Reverse Causal Inference on Panel Data via Generalized Synthetic Control.** *NeurIPS Workshop on Causal Dynamic System*, 2022.

[30] TS A. Jambulapati, H. Hasson, **Y. Park**, Y. Wang. **Testing Causality of High-Dimensional Data.** Available at ArXiv, 2022.

[31] TS **Y. Park**. **On the Explainability of Deep Forecasting Models.** *Amazon Machine Learning Conference (AMLC)*, 2021.

[32] OPT TS Y. Lu, **Y. Park†**, L. Cheng, Y. Wang, D. Foster. **Variance Reduced Training with Stratified Sampling for Forecasting Models.** *International Conference on Machine Learning (ICML)*, 2021.

[33] OPT RL **Y. Park**, R. Rossi, Z. Wen, G. Wu, H. Zhao. **Structured Policy Iteration for Linear Quadratic Regulator.** *International Conference on Machine Learning (ICML)*, 2020.

[34] OPT J. Kim, **Y. Park**, J. Fox, S. Boyd, W. Dally. **Optimal Operation of a Plug-in Hybrid Vehicle with Battery Thermal and Degradation Model.** *American Control Conference (ACC)*, 2020.

[35] OPT **Y. Park**, S. Dhar, S. Boyd, M. Shah. **Variable Metric Proximal Gradient Method with Diagonal Barzilai-Borwien Stepsize.** *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020.

[36] OPT **Y. Park**, E. K. Ryu. **Linear Convergence of Cyclic SAGA.** *Optimization Letters*, 2020.

[37] OPT RL **Y. Park**, K. Mahadik, R. Rossi, G. Wu, H. Zhao. **Linear Quadratic Regulator for Resource-Efficient Cloud Services.** *ACM Symposium on Cloud Computing (SOCC)*, 2019.

[38] OPT ML **Y. Park**, D. Hallac, S. Boyd, J. Leskovec. **Learning the Network Structure of Heterogeneous Data via Pairwise Exponential Markov Random Fields.** *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[39] OPT ML D. Hallac, **Y. Park**, S. Boyd, J. Leskovec. **Inferring Time Varying Networks via Graphical Lasso.** *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2017.

†Corresponding author

Tutorial & Teaching Experience

Conference Tutorials & Workshops:

- **AAAI 2026:** Algorithms and Systems for Efficient Inference in Generative AI (with R. Saha, Y. Wang)
- **IJCAI 2025:** Scaling LLM Training: Efficient Pre-training & Fine-tuning (with T. Yu, L. Lausen)

- **KDD 2024:** Inference Optimization of Foundation Models (with K. Budhathoki, J. Kübler, Y. Wang)
- **KDD 2023:** Training Foundation Models on Emerging AI Chips (with A. Muhamed, J. Huan)
- **IEEE Big Data 2022:** Deep Time Series Forecasting (with G. Gupta, J. Huan)

Workshop Organization:

- **KDD 2025:** Workshop on Inference Optimization for Generative AI, Toronto, Canada
- **KDD 2022:** Workshop on Mining and Learning from Time Series – Deep Forecasting, Washington DC

Stanford University:

- Teaching Assistant: Convex Optimization I (Instructor: Stephen Boyd), 2018
- Teaching Assistant: Convex Optimization II (Instructor: John Duchi), 2017

Honors & Awards

- Top 10 Most Read Blog, Amazon Science, Dec. 2022
- Top 3 Most Read Paper, Amazon Science, Jun. 2022
- Best Presenter Award in AI Session, Hyundai Global Forum, 2018
- Kwanjeong Graduate Fellowship ($110,000 over 2 years), 2013–2015
- Fulbright Graduate Fellowship (Declined), 2013
- National Science and Engineering Scholarship, KOSAF, 2006–2009

Invited Talks (Selected)

2025	King Abdullah University of Science and Technology (KAUST), Saudi Arabia
2023	University of California, San Diego
2022	Amazon Machine Learning Conference (AMLC), Virtual
2022	AWS AI Labs, Santa Clara
2021	Amazon Machine Learning Conference, Virtual
2020	Seoul National University, South Korea
2020	Amazon Web Service (AWS) AI, Palo Alto
2020	Facebook AI Research (FAIR), Menlo Park
2020	Rakuten Research, San Mateo
2019	Adobe Research, San Jose
2019	Hyundai AI Labs, Seoul, Korea
2018	Hyundai Global Forum, San Diego
2017	Kakao Brain, Bundang, Korea
2017	Bosch AI, Palo Alto

Patents

- Model Explainability Insight for Time Series Forecasting (US Patent, 2021)
- System and Method for Resource Scaling for Efficient Resource Management (US Patent, 2020)

Open-Source Contributions

- **NxDT** (Neuron Distributed Training): Foundation model training framework for AWS Trainium
  https://awsdocs-neuron.readthedocs-hosted.com/
- **GluonTS**: Probabilistic time series modeling in Python (1.5K+ GitHub stars)
  https://github.com/awslabs/gluon-ts
- **SnapVX**: Convex optimization solver for problems on graphs
  http://snap.stanford.edu/snapvx/
- **TVGL**: Time-varying graphical lasso for network inference

https://github.com/davidhallac/TVGL

- **PEMRF**: Graphical structure inference via pairwise exponential MRF
https://github.com/youngsuk0723/PE-MRF-Code

Academic Collaborators

Stanford: Stephen P. Boyd, Jure Leskovec, Tsachy Weissman, Michael Saunders
Other Academia: Ernest K. Ryu (Seoul National U), Mingyi Hong (U Minnesota), Volkan Cevher (EPFL), Yuxiang Wang (UCSD), Hongseok Namkoong (Columbia)
Industry Research: George Karypis (AWS), Luke Huan (AWS), Yida Wang (AWS), Yuyang Wang (AWS), Dean Foster (Amazon), Branislav Kveton (Adobe Research), Zheng Wen (Google DeepMind), Suju Rajan (LinkedIn), Mohak Shah (LG Electronics)

Professional Service

Conference Reviewing: NeurIPS, ICML, ICLR, AISTATS, KDD
Journal Reviewing: JMLR, TMLR, IEEE TPAMI, SIAM Journal on Mathematics of Data Science

Research Mentorship

- **Jason Yoo**, University of British Columbia
- **Wenlong Deng**, University of British Columbia
- **Jiin Woo**, CMU Computer Science → AWS AI
- **Ahmed Khaled Ragab**, Princeton EECS → Google DeepMind
- **Song Bian**, UW-Madison Computer Science
- **Jaihoon Kim**, KAIST Computer Science
- **Albert Tseng**, Cornell Computer Science → OpenAI
- **Kaan Ozkara**, UCLA Electrical Engineering → AWS AI
- **Licong Lin**, UC Berkeley Statistics
- **Chung Yiu Yau**, Chinese University of Hong Kong Computer Science
- **Hongyi Liu**, Rice Computer Science → AWS AI
- **Tanmay Gautam**, UC Berkeley EECS → Microsoft
- **Shingen Sun**, Northwestern Applied Mathematics
- **Tao Yu**, Cornell Computer Science → AWS AI
- **Charlie Marx**, Stanford Computer Science → Hedge fund
- **Yuhao Ding**, UC Berkeley Operations Research → Hedge fund
- **Linbo Liu**, UCSD Applied Mathematics → AWS AI
- **Sanae Lotfi**, NYU Statistics → Meta FAIR
- **Jiayao Zhang**, UPenn Applied Mathematics → Hedge fund
- **Luca Masserano**, CMU Statistics → Meta
- **Xiyuan Zhang**, UCSD Computer Science → AWS AI
- **Arun Jambulapati**, Stanford Mathematics → Postdoc
- **Kelvin Kan**, Emory Mathematics → Postdoc
- **Shantnu Gupta**, CMU Computer Science
- **Taeho Yoon**, Seoul National University Mathematics → Postdoc
- **Yucheng Lu**, Cornell Computer Science → TogetherAI

- **Xiaoyong Jin**, UCSB Computer Science → AWS AI