ROBERT DYRO

robert.dyro@gmail.com | (310) 694-1753 | https://robertdyro.com

I am interested in cutting-edge computational engineering research. My current focus is on computational frameworks that accelerate research iteration and model development. I am passionate about exploring new, high-impact technologies. I thrive in dynamic, collaborative and results-driven environments.

EDUCATION

Stanford University — Stanford, CA
PhD, Robotics, GPA 3.93 — 2020 - 2024
MS, Aeronautics & Astronautics Engineering, GPA 3.89 — 2018 - 2020

University of California, Los Angeles — Los Angeles, CA
BS, Aerospace Engineering, Minor in Philosophy, GPA 3.94, Summa Cum Laude — 2014 - 2018

RELEVANT COURSEWORK

Convex Optimization ■ RL ■ Meta-Learning ■ Graph ML ■ Trustworthy & Explainable ML ■ ML under Distribution Shift

EXPERIENCE

Software Engineer, JAX External at Google — Mountain View, CA
- JAX development for external researchers and industry partners — 2024 - present
- Built a minimalistic LLM serving framework (`jax-llm-examples`) for minimum-latency model serving in JAX
- LLM inference optimization in multi-node inference deployments (TPU & GPU) for dense and MoE models
- Maintaining and developing cutting-edge flash attention and ragged dot (gmm) kernels for training in JAX-Pallas for TPU
- Working on new generation TPU hardware, training and inference optimizations for LLMs
- Working with open-source LLM training team on training performance - kernels and model sharding strategies

Graduate Student, Autonomous Systems Laboratory (ASL) at Stanford University — Stanford, CA

Stress Testing Autonomous Vehicles via Counterfactual Editing of Trained Behavior Models — 2023
- Extracting learned behavior distribution for realistic counterfactual generation via efficient and scalable Hessian sketching

Optimization-based Online Intent Inference in Autonomous Driving — 2022
- Developed a real-time, structured behavior inference method for online behavior identification in autonomous driving

Second-Order Sensitivity Analysis for Bilevel Optimization — 2021
- 2nd order sensitivity analysis of optimization, enabling much faster optimization of bilevel/inverse/sensitivity problems

Control under Arbitrary Uncertainty using Particle Model Predictive Control — 2020
- Implemented and experimentally evaluated consensus control particle MPC for control under arbitrary uncertainty

Convex Last-layer Meta-learning for Behavior & Physics-based Modeling — 2019
- Incorporated constraints into the meta-learning model for structured learning to allow adding a priori modeling knowledge

PhD Intern, Cruise — San Francisco, CA
Machine Learning Acceleration - Architecture Optimization - Zero-Shot Neural Architecture Search — June - December 2022

Research Intern, Toyota Research Institute — Los Altos, CA
Intelligent Driver Behavior Modeling using Human Interpretable Rules — June - September 2020
- Embedded human logic within path planning via Signal Temporal Logic (STL) to capture human-interpretable specifications

TECHNICAL EXPERIENCE

Projects:
- *tune-jax* - automatic kernel tuning library for JAX and Pallas
- *torch2jax* - zero-overhead PyTorch computation wrapping for JAX computation graph under JIT and autodifferentiation
- neural architecture search for the most generalizing graph neural network via GraphGym
- custom quadratic program (QP) solver in CUDA
- experimental dynamic graph autodifferentiation library for full sparse 1st & 2nd order matrix algebra differentiation

Software Skills:
Python, C++, C, Julia, Matlab ■ JAX, PyTorch ■ Linux, HPC, Slurm, CUDA, Google Cloud

MISC

Philosophy Minor, UCLA ■ LA Marathon ■ Amateur Radio License ■ PADI Assistant Instructor