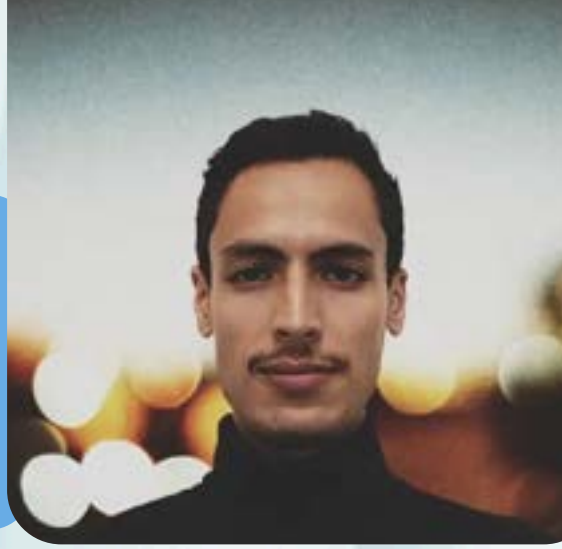
JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING | Electrical and Computer Engineering



SEMINAR

SOUFIANE HAYOU

ASSISTANT PROFESSOR
JOHNS HOPKINS UNIVERSITY, DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS

EFFICIENT FINETUNING OF LARGE LANGUAGE MODELS VIA LARGE-WIDTH ANALYSIS

OCTOBER 2, 2025

3:00-4:15 P.M.

REMSEN 1

Abstract: Finetuning Large Language Models (LLMs) enhances their performance on downstream tasks — a desirable outcome if the model is used for a specific task. Parameter-efficient finetuning methods such as LoRA (Low-Rank Adaptation) are popular because they allow finetuning large models with relatively low cost. When using LoRA, two hyperparameters critically shape learning: learning rates and initialization. In this talk, I'll present two results. First, we prove and demonstrate that the two "zero-product" initializations (A random/B=0 vs. B random/A=0) are not equivalent: initializing B=0, A random permits larger stable learning rates and yields better performance, with an infinite-width stability analysis explaining the gap and LLM experiments confirming it. Second, LoRA+ shows that using the same learning rate for the A and B matrices is suboptimal at large width; a simple asymmetric LR scheme yields more efficient feature learning and delivers consistent accuracy gains and up to ~2× faster convergence at the same compute. Finally, I will distill these insights into practical defaults.

Bio: Soufiane Hayou is currently an assistant professor at Johns Hopkins in the department of Applied Mathematics and Statistics. He is also a member of the Data Science and AI Institute. Previously, he was a postdoc researcher at Simons Institute, UC Berkeley, and a visiting assistant professor of mathematics at the National University of Singapore. He obtained his PhD in statistics and machine learning in 2021 from the University of Oxford, and graduated from Ecole Polytechnique in 2018 before joining Oxford. His research is mainly focused on the theory and practice of learning at scale: theoretical analysis of large scale neural networks with the goal of obtaining principled methods for training/finetuning. Topics include depth scaling (Stable ResNet), hyperparameter transfer (Depth-muP parametrization), efficient finetuning (LoRA+), etc.