Dynamics of neural scaling laws in random feature regression with powerlaw-distributed kernel eigenvalues

J. Kramp[1,2] J. Lindner[1,2] and M. Helias[1,3]

[1] Institute for Advanced Simulation (IAS-6), Computational and Systems Neuroscience, Jülich Research Centre, Jülich, Germany, [2] RWTH Aachen University, Aachen, Germany, [3] Department of Physics, RWTH Aachen University, Aachen, Germany

Training large neural networks exposes neural scaling laws for the generalization error, which points to a universal behavior across network architectures of learning in high dimensions. It was also shown that this effect persists in the limit of highly overparametrized networks as well as the Neural network Gaussian process limit. We here develop a principled understanding of the typical behavior of generalization in Neural Network Gaussian process regression dynamics. We derive a dynamical mean-field theory that captures the typical case learning dynamics: This allows us to unify multiple existing regimes of learning studied in the current literature, namely Bayesian inference on Gaussian processes, gradient flow with or without weight-decay, and stochastic Langevin training dynamics. Employing tools from statistical physics, the unified framework we derive in either of these cases yields an effective description of the high-dimensional microscopic behavior of networks dynamics in terms of lower dimensional order parameters. We show that collective training dynamics may be separated into the dynamics of N independent eigenmodes, whose evolution equations are only coupled through collective response functions and a common statistics of an effective, independent noise. Our approach allows us to quantitatively explain the dynamics of the generalization error by linking spectral and dynamical properties of learning on data with power law spectra, including phenomena such as neural scaling laws and the effect of early stopping.

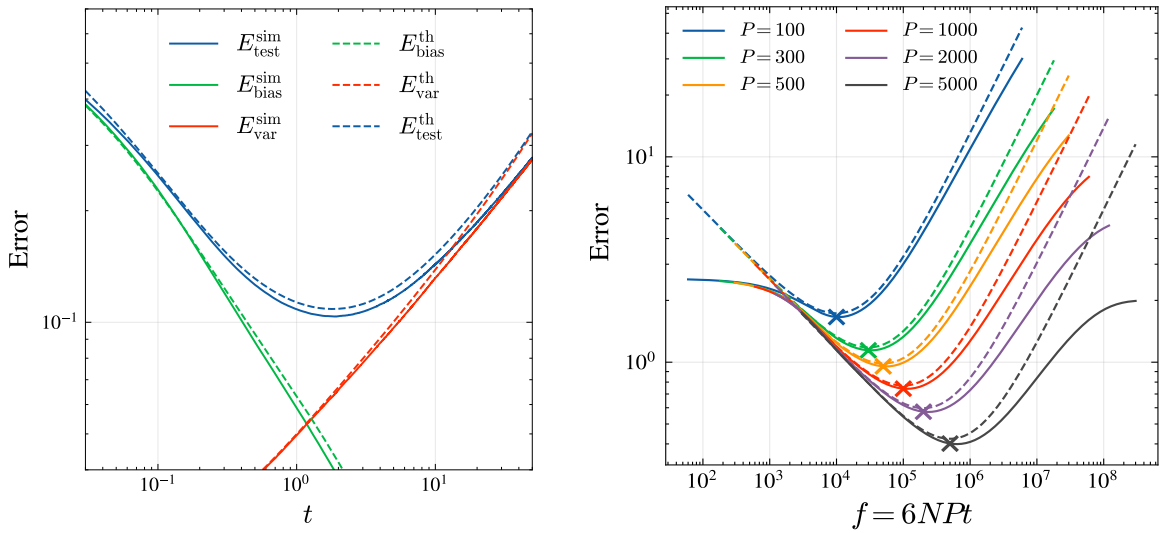

Figure 1: Left: Bias-variance decomposition of generalization error as a function of time. Right: Pareto frontier of the generalization error as a function of compute f for different sample sizes P.

[1] J. Kramp et al., arXiv preprint, arXiv:2602.23039, 2026..