ANASTASIS KRATSIOS, McMaster University and the Vector Institute
Pathwise Generalization bounds for Transformers

We derive non-asymptotic statistical guarantees in this setting through bounds on the *generalization* of a transformer network at a future-time t, given that it has been trained using $N \leq t$ observations from a single perturbed trajectory of a Markov process. Under the assumption that the Markov process satisfies a log-Sobolev inequality, we obtain a generalization bound which effectively converges at the rate of $\mathcal{O}(1/\sqrt{N})$. Our bound depends explicitly on the activation function (Swish, GeLU, or $\tanh$ are considered), the number of self-attention heads, depth, width, and norm-bounds defining the transformer architecture.

Joint work: Blanka Horvath and Yannick Limmer (Oxford Math), Xuwei Yang (McMaster), and Raeid Saqur (U. Toronto and Princeton).