Sampling at intermediate temperatures is optimal for training large language models in protein structure prediction

Alessandro Zambon[1,2], Luca Ghiringhelli[1], Francesca Caruso[3], Guido Tiana[1,2], Riccardo Zecchina[3]

[1]Department of Physics, Università degli Studi di Milano, Milan, Italy, [2]INFN, Milan, Italy, [3]Department of Computing Sciences and Bocconi Institute for Data Science and Analytics (BIDSA), Bocconi University, Milan, Italy

Sampling the parameter space of artificial neural networks according to a Boltzmann distribution is a powerful approach for studying the geometry and structure of low-loss solutions, providing a principled alternative to conventional deterministic loss minimization during network training. In this framework, configurations of network parameters are explored probabilistically, enabling a more comprehensive characterization of the loss landscape. However, although exact sampling methods such as hybrid Monte Carlo are theoretically sound and asymptotically exact, they quickly become computationally intractable for large-scale, realistic datasets, since they require repeated evaluations of full-batch gradients, resulting in prohibitive computational costs.

To address this issue, we propose a pseudo-Langevin dynamics scheme that enables efficient approximate Boltzmann sampling for neural networks trained on large datasets. Our approach leverages mini-batch gradients in a controlled and theoretically informed way, exploiting the statistical structure of the noise they introduce. By carefully tuning fictitious masses and friction coefficients, we ensure that the resulting stochastic dynamics closely approximate the desired equilibrium distribution, all the while remaining compatible with the computational efficiency required for modern, large-scale learning setups.

Using this method, we explore the parameter space of transformer architectures trained on protein sequence data within a statistical mechanics framework. By sampling the loss landscape across a range of effective temperatures, we characterize the structure of the low-loss manifold and investigate the mechanisms that underlie the strong empirical performance of transformers in protein structure prediction tasks. In contrast to feedforward networks, we find no evidence of a first-order phase transition in the loss landscape of transformers. Instead, a broad regime of intermediate temperatures exists that exhibits favorable learning properties.

Within this regime, we find that, if the embedding dimension is chosen correctly, the parameters across most layers remain highly conserved. This observation enables us to propose a practical criterion for identifying optimal embedding sizes. Finally, we demonstrate that, at higher temperatures and larger embedding dimensions, the attention matrices increasingly predict protein contact maps, even beyond the point at which standard learning performance is optimized. This highlights a trade-off between interpretability and accuracy and suggests promising avenues for future methodological and theoretical developments in this field.