Proceedings of Machine Learning Research vol :1–1, 2026 39th Annual Conference on Learning Theory

Universality of high-dimensional scaling limits of stochastic gradient descent

Reza Gheissari GHEISSARI@NORTHWESTERN.EDU
Department of Mathematics, Northwestern University

Aukosh Jagannath A.JAGANNATH@UWATERLOO.CA
Department of Statistics and Actuarial Science, Department of Applied Mathematics, and Cheriton School of Computer Science, University of Waterloo

Editors: Steve Hanneke and Tor Lattimore

Abstract

We consider statistical tasks in high dimensions whose loss depends on the data only through its projection into a fixed-dimensional subspace spanned by the parameter vectors and certain ground truth vectors. This includes classifying mixture distributions with cross-entropy loss with one and two-layer networks, and learning single and multi-index models with one and two-layer networks. When the data is drawn from an isotropic Gaussian mixture distribution, it is known that the evolution of a finite family of summary statistics under online stochastic gradient descent (SGD) converges to an autonomous ordinary differential equation (ODE), as the dimension and sample size go to infinity proportionately and the step size goes to zero commensurately. Furthermore their fluctuations about fixed points are known to converge to stochastic differential equations. Results of this type are called high-dimensional scaling limits.

In this paper, we seek to understand to what extent these high-dimensional scaling limits are "universal". That is, to what extent these scaling limits are agnostic to the specific properties of the data distribution beyond the first two moments. By contrast, it is important to note that the classical asymptotic limit theory (when dimension is fixed and step size goes to zero) is *not* universal as the population loss can depend on high moments. Our main result is that in the high-dimensional setting, the limits are indeed universal.

To do this, we extend the "effective dynamics" framework of Ben Arous and the two authors (2022). That framework does not explicitly require Gaussianity. However, the central assumption of that work is "asymptotic closability" which (loosely speaking) says that certain expectations of gradients of the loss are close to functions of a fixed family of summary statistics *uniformly* over the parameter space. This assumption, however, fails in many models of interest when the data is non-Gaussian. In this paper, we show that, even without Gaussianity, this assumption holds *locally* in the region of the parameter space explored by SGD provided its initialization is suitably coordinate delocalized. The key technical contribution is to show that if SGD is initialized in a delocalized region, it remains delocalized for long enough time-scales.

We complement this by proving two corresponding non-universality results. We provide a simple example where the ODE limits are non-universal if the initialization is coordinate aligned. We also show that the stochastic differential equation limits arising as fluctuations of the summary statistics around their ODE's fixed points are generally not universal, even with optimally coordinate delocalized initialization.[1]

Keywords: stochastic gradient descent, high-dimensional limits, universality, Gaussian equivalence, multi-index models, classification

1. Extended abstract. Full version appears as [arXiv 2512.13634, v2]