Universality of empirical risk minimization

Andrea Montanari
Department of Electrical Engineering and Department of Statistics, Stanford University

Basil Saeed
Department of Electrical Engineering, Stanford University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract
Consider\(^1\) supervised learning from i.i.d. samples \(\{(y_i, x_i)\}_{i \leq n}\) where \(x_i \in \mathbb{R}^p\) are feature vectors and \(y_i \in \mathbb{R}\) are labels. We study empirical risk minimization over a class of functions that are parameterized by \(K = O(1)\) vectors \(\theta_1, \ldots, \theta_k \in \mathbb{R}^p\), and prove universality results both for the training and test error. Namely, under the proportional asymptotics \(n, p \to \infty\), with \(n/p = \Theta(1)\), we prove that the training error depends on the random features distribution only through its mean and covariance structure. We also prove that the minimum test error over near-empirical risk minimizers enjoys similar universality properties. Furthermore, we give conditions guaranteeing universality of the test error of the empirical risk minimizer that can be checked in a “Gaussian equivalent” model where the features are replaced with Gaussian features of the same (asymptotic) mean and covariance. In particular, the asymptotics of the train and test error can be computed —to leading order— under a simpler model in which the feature vectors \(x_i\) are replaced by Gaussian vectors \(g_i\) with the same mean and covariance.

Earlier universality results were limited to strongly convex learning procedures, or to feature vectors \(x_i\) with independent entries. Our main results hold for non-convex procedures and feature vectors with dependent entries as long as they have asymptotically Gaussian projections along vectors \(\theta\) whose \(\ell_2\) norm is “well-spread out” among its coordinates. We give examples showing that generally, for universality to hold, one needs to constrain the minimization of the empirical risk to this set of well-spread out vectors. Furthermore, we give examples of conditions under which the minimum over this restricted space converges to the unrestricted minimum.

Our distributional assumptions are general enough to include feature vectors \(x_i\) that are produced by randomized featurization maps. In particular we explicitly check the assumptions for certain random features models (computing the output of a one-layer neural network with random weights) and neural tangent models (first-order Taylor approximation of two-layer networks).

Keywords: High-dimensional statistics, neural networks/deep learning, statistical physics

Acknowledgements
This work was supported by the NSF through award DMS-2031883, the Simons Foundation through Award 814639 for the Collaboration on the Theoretical Foundations of Deep Learning, the NSF grant CCF-2006489, the ONR grant N00014-18-1-2729, and an NSF GRFP award.

---

1. Extended abstract. Full version appears on arXiv as 2202.08832

© 2022 A. Montanari & B. Saeed.
References


