Universality of empirical risk minimization

Andrea Montanari MONTANARI@STANFORD.EDU
Department of Electrical Engineering and Department of Statistics, Stanford University

Basil Saeed BSAEED@STANFORD.EDU
Department of Electrical Engineering, Stanford University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Consider[1] supervised learning from i.i.d. samples $\{(y_i, \boldsymbol{x}_i)\}_{i\le n}$ where $\boldsymbol{x}_i \in \mathbb{R}^p$ 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 $\mathsf{k} = O(1)$ vectors $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{\mathsf{k}} \in \mathbb{R}^p$, 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 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 $\boldsymbol{x}_i$ are replaced by Gaussian vectors $\boldsymbol{g}_i$ with the same mean and covariance.

Earlier universality results were limited to strongly convex learning procedures, or to feature vectors $\boldsymbol{x}_i$ 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 $\boldsymbol{\theta}$ whose ℓ_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 unristricted minimum.

Our distributional assumptions are general enough to include feature vectors $\boldsymbol{x}_i$ 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

References

Benjamin Aubin, Florent Krzakala, Yue Lu, and Lenka Zdeborová. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33:12199–12210, 2020.

Mohsen Bayati and Andrea Montanari. The LASSO risk for Gaussian matrices. *IEEE Trans. on Inform. Theory*, 58:1997–2017, 2012.

Michael Celentano and Andrea Montanari. Fundamental barriers to high-dimensional regression with convex penalties. *Annals of Statistics*, 2022.

Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general Gaussian designs with applications to hypothesis testing. *arXiv:2007.13716*, 2020.

Sourav Chatterjee. A generalization of the lindeberg principle. *The Annals of Probability*, 34(6): 2061–2076, 2006.

Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*, volume 2. Springer, 2011.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv:1911.05822*, 2019.

David Donoho and Andrea Montanari. High dimensional robust M-estimation: asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3):935–969, Dec 2016.

Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175, 2018.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020a.

Sebastian Goldt, Galen Reeves, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with two-layer neural networks. *arXiv:2006.14709*, 2020b.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv:1903.08560*, 2019.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv:2009.07669*, 2020.

Charles R Johnson. *Matrix theory and applications*, volume 40. American Mathematical Soc., 1990.

Satish Babu Korada and Andrea Montanari. Applications of the lindeberg principle in communications and statistical learning. *IEEE transactions on information theory*, 57(4):2440–2450, 2011.

Jarl Waldemar Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225, 1922.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

Andrea Montanari and Phan-Minh Nguyen. Universality of the elastic net error. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2338–2342. IEEE, 2017.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv:1911.01544*, 2019.

Samet Oymak and Joel A Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 2018.

Ashkan Panahi and Babak Hassibi. A universal analysis of large-scale regularized least squares solutions. *Advances in Neural Information Processing Systems*, 30, 2017.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Pragya Sur, Yuxin Chen, and Emmanuel J Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability theory and related fields*, 175(1):487–558, 2019.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1683–1709, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Thrampoulidis15.html.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.