Proceedings of Machine Learning Research vol 178:1–3, 2022 35th Annual Conference on Learning Theory

High-Dimensional Projection Pursuit: Outer Bounds and Applications to Interpolation in Neural Networks

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

Kangjie Zhou KANGJIE@STANFORD.EDU
Department of Statistics, Stanford University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Given[1] a cloud of n data points in $\mathbb{R}^d$, consider all projections onto m-dimensional subspaces of $\mathbb{R}^d$ and, for each such projection, the empirical distribution of the projected points. What does this collection of probability distributions look like when n, d grow large? We consider this question under the null model in which the points are i.i.d. standard Gaussian vectors, focusing on the asymptotic regime in which $n, d \to \infty$, with $n/d \to \alpha \in (0, \infty)$, while m is fixed. Denoting by $\mathscr{F}_{m,\alpha}$ the set of probability distributions in $\mathbb{R}^m$ that arise as low-dimensional projections in this limit, we establish several new results on this model:

Wasserstein radius for $m = 1$. Denoting by $W_2(P_1, P_2)$ the second Wasserstein distance between probability measures P_1 and P_2, we prove that $\sup\{W_2(P, \mathsf{N}(0,1)) : P \in \mathscr{F}_{1,\alpha}\} = 1/\sqrt{\alpha}$.

KL-Wasserstein outer bound. We show that, for any m, $\mathscr{F}_{m,\alpha}$ is contained in a W_2 neighborhood of the set of distributions P such that $D_{\rm KL}(P \| \mathsf{N}(\mathbf{0}, \boldsymbol{I}_m)) \le Cm\alpha^{-1}(1 \vee \log \alpha)$, with $D_{\rm KL}$ the Kullback-Leibler divergence.

Information dimension bound. Denoting by $\underline{d}(P)$ the lower information dimension of P, we prove that $\mathscr{F}_{m,\alpha}$ is contained in $\{P : \underline{d}(P) \ge m(1 - 1/\alpha)\}$ for $\alpha > 1$.

The previous question has application to unsupervised learning methods, such as projection pursuit and independent component analysis. We introduce a version of the same problem that is relevant for supervised learning, where the labels depend on k-dimensional projections of the covariates through a link function φ, and present the following results:

General ERM asymptotics. We consider a class of empirical risk minimization problems over functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form $f(\boldsymbol{x}) = h(\boldsymbol{W}^\top \boldsymbol{x})$, and show that the asymptotics of the minimum empirical risk can be expressed in terms of the feasibility set $\mathscr{F}^{\varphi}_{m,\alpha}$.

Wasserstein bound for $m = 1$. We prove an outer bound on $\mathscr{F}^{\varphi}_{1,\alpha}$ for general $k = O(1)$, which generalizes the Wasserstein radius result obtained in the unsupervised setting. In fact, this outer bound characterizes the maximum W_2 distance between the empirical distribution of one-dimensional projections and the expected distribution.

Interpolation for two-layer networks. As a corollary to the previous result, we prove that a neural network with two-layers and m hidden neurons can separate n data points in d dimensions with margin κ only if $md \ge C\kappa^2 n$. Earlier bounds only required $md \ge Cn/\log(d/\kappa)$.

Margin distributions for linear classifier. We demonstrate the tightness of our W_2 bound by deriving the asymptotic distribution of the margins in linear max-margin classification.

1. Extended abstract. Full version appears as [arXiv:2206.06526, v1]

References

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Peter J Bickel, Gil Kur, and Boaz Nadler. Projection pursuit in high dimensions. *Proceedings of the National Academy of Sciences*, 115(37):9151–9156, 2018.

Gilles Blanchard, Motoaki Kawanabe, Masashi Sugiyama, Vladimir Spokoiny, Klaus-Robert Müller, and Sam Roweis. In search of non-gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, 7(2), 2006.

Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48(1):27–42, 2020.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer Berlin Heidelberg, 2010. doi: 10.1007/978-3-642-03311-7. URL https://doi.org/10.1007%2F978-3-642-03311-7.

Persi Diaconis and David Freedman. Asymptotics of graphical projection pursuit. *The annals of statistics*, pages 793–815, 1984.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Peter Eichelsbacher and Uwe Schmock. *Large deviations of products of empirical measures and U-Empirical measures in strong topologies*. Univ. Bielefeld, Sonderforschungsbereich 343, Diskrete Strukturen in der Math., 1996.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Jerome H Friedman and John W Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9):881–890, 1974.

Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding blind deconvolution algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2354–2367, 2011.

Nicola Loperfido. Skewness-based projection pursuit: A computational approach. *Computational Statistics & Data Analysis*, 120:42–57, 2018.

Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335, 2021.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *arXiv preprint arXiv:2007.12826*, 2020.

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 preprint arXiv:1911.01544*, 2019.

Alfréd Rényi. On the dimension and entropy of probability distributions. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(1-2):193–215, 1959.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

Hiroaki Sasaki, Gang Niu, and Masashi Sugiyama. Non-gaussian component analysis with log-density gradient estimation. In *Artificial Intelligence and Statistics*, pages 1177–1185. PMLR, 2016.

Michel Talagrand. Transportation cost for gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6(3):587–600, 1996.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. *Proceedings of Machine Learning Research*, 40:1683–1709, 2015.

Aad W Van Der Vaart and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

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

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Ran Wang, Xinyi Wang, and Liming Wu. Sanov's theorem in the Wasserstein distance: a necessary and sufficient condition. *Statistics & Probability Letters*, 80(5-6):505–512, 2010. ISSN 0167-7152. doi: 10.1016/j.spl.2009.12.003.