TU Delft

Delft University of Technology

A brief prehistory of double descent

Loog, Marco; Viering, Tom; Mey, Alexander; Krijthe, Jesse H.; Tax, David M. J.

DOI
10.1073/pnas.2001875117

Publication date
2020

Document Version
Final published version

Published in
Proceedings of the National Academy of Sciences of the United States of America

Citation (APA)
Loog, M., Viering, T., Mey, A., Krijthe, J. H., & Tax, D. M. J. (2020). A brief prehistory of double descent. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(20), 10625-10626. https://doi.org/10.1073/pnas.2001875117

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

A brief prehistory of double descent

Marco Loog[a,b,1], Tom Viering[a], Alexander Mey[a], Jesse H. Krijthe[a], and David M. J. Tax[a]

In their thought-provoking paper, Belkin et al. (1) illustrate and discuss the shape of risk curves in the context of modern high-complexity learners. Given a fixed training sample size n, such curves show the risk of a learner as a function of some (approximate) measure of its complexity N. With N the number of features, these curves are also referred to as feature curves. A salient observation in ref. 1 is that these curves can display what they call double descent: With increasing N, the risk initially decreases, attains a minimum, and then increases until N equals n, where the training data are fitted perfectly. Increasing N even further, the risk decreases a second and final time, creating a peak at $N = n$. This twofold descent may come as a surprise, but as opposed to what ref. 1 reports, it has not been overlooked historically. Our letter draws attention to some original earlier findings of interest to contemporary machine learning.

Already in 1989, using artificial data, Vallet et al. (2) experimentally demonstrated double descent for learning curves of classifiers trained through minimum norm linear regression (MNLR; see ref. 3)—termed the pseudo-inverse solution in ref. 2. In learning curves the risk is displayed as a function of n, as opposed to N for risk curves. What intuitively matters in learning behavior, however, is the sample size relative to the measure of complexity. This idea is made explicit in various physics papers on learning (e.g., refs. 2, 4, and 5), where the risk is often plotted against $\alpha = \frac{n}{N}$. A first theoretical result on double descent, indeed using such α, is given by Opper et al. (4). They prove that in particular settings, for N going to infinity, the pseudo-inverse solution improves as soon as one moves away from the peak at $\alpha = 1$.

Employing a so-called pseudo-Fisher linear discriminant (PFLD, equivalent to MNLR), Duin (6) was the first to show feature curves on real-world data quite similar to the double-descent curves in ref. 1. Compare, for instance, figure 2 in ref. 1 with figures 6 and 7 in ref. 6. Skurichina and Duin (7) demonstrate experimentally that increasing PFLD's complexity simply by adding random features can improve performance when $N = n$ (i.e., $\alpha = 1$). The benefit of some form of regularization has been shown already in ref. 2. For semisupervised PFLD, Krijthe and Loog (8) demonstrate that unlabeled data can regularize but also worsen the peak in the curve. Their work builds on the original analysis of double descent for the supervised PFLD by Raudys and Duin (9).

Interestingly, results from refs. 4–7 suggest that some losses may not exhibit double descent in the first place. In refs. 6 and 7, the linear support vector machine (SVM) shows regular monotonic behavior. Analytic results from refs. 4 and 5 show the same for the so-called perceptron of optimal (or maximal) stability, which is closely related to the SVM (5).

The findings in ref. 1 go, significantly, beyond those for the MNLR. Also shown, for instance, is double descent for two-layer neural networks and random forests. Combining this with observations such as those from Loog et al. (10), which show striking multiple-descent learning curves (even in the underparameterized regime), the need to further our understanding of such rudimentary learning behavior is evident.

1 M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 15849–15854 (2019).
2 F. Vallet, J.-G. Cailton, Ph. Refregier, Linear and nonlinear extension of the pseudo-inverse solution for learning Boolean functions. *Europhys. Lett.* **9**, 315–320 (1989).
3 R. Penrose, On best approximate solutions of linear matrix equations. *Math. Proc. Cambridge Philos. Soc.* **52**, 17–19 (1956).
4 M. Opper, W. Kinzel, J. Kleinz, R. Nehl, On the ability of the optimal perceptron to generalise. *J. Phys. A Math. Gen.* **23**, L581–L586 (1990).
5 T. L. H. Watkin, A. Rau, M. Biehl, The statistical mechanics of learning a rule. *Rev. Mod. Phys.* **65**, 499 (1993).

[a]Pattern Recognition Laboratory, Delft University of Technology, 2628 CD Delft, The Netherlands; and [b]Department of Computer Science, University of Copenhagen, 1165 Copenhagen, Denmark

Author contributions: M.L., T.V., and A.M. performed the literature research; M.L. wrote the paper; and M.L., T.V., A.M., J.H.K., and D.M.J.T. provided comments, suggestions, and the core thoughts and arguments presented.

The authors declare no competing interest.

Published under the PNAS license.

[1]To whom correspondence may be addressed. Email: m.loog@tudelft.nl.

First published May 5, 2020.

Downloaded at Technische Universiteit Delft on December 14, 2020

6 R. P. W. Duin, "Classifiers in almost empty spaces" in *Proceedings of the 15th International Conference on Pattern Recognition* (IEEE, 2000), vol. 2, pp. 1–7.
7 M. Skurichina, R. P. W. Duin, "Regularization by adding redundant features" in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (Springer, 1998), pp. 564–572.
8 J. H. Krijthe, M. Loog, "The peaking phenomenon in semi-supervised learning" in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* (Springer, 2016), pp. 299–309.
9 Š. Raudys, R. P. W. Duin, Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognit. Lett.* **19**, 385–392 (1998).
10 M. Loog, T. Viering, A. Mey, "Minimizers of the empirical risk and risk monotonicity" in *Advances in Neural Information Processing Systems* (MIT Press, 2019), pp. 7476–7485.

Downloaded at Technische Universiteit Delft on December 14, 2020