Notes of a Numerical Analyst

Double Descent

NICK TREFETHEN FRS

Take the function $f(x) = \tanh(10x)$ and sample it at 51 equispaced points from -1 to 1. Then fit this data in the least-squares sense by a degree n polynomial p_n, $0 \le n \le 100$. What's the ∞-norm error $\|f - p_n\|$ over $[-1, 1]$ as a function of n? Figure 1 shows the surprising answer.

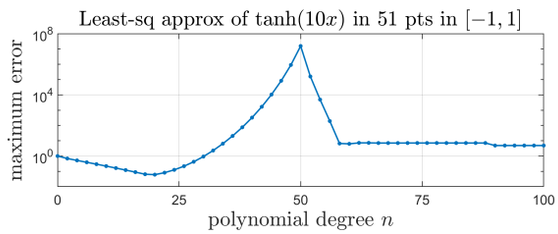

Figure 1. Overdetermined ($n < 50$) and underdetermined ($n > 50$) polynomial least-squares fitting illustrates double descent.

The first half of the figure goes back to 1901. At first, as n increases from 0, p_n begins to fit the data. But then the *Runge phenomenon* of overfitting sets in, causing huge oscillations between sample points for $x \approx \pm 1$. For this choice of f and 51 points, the peak is reached at degree $n = 50$, where least-squares fitting becomes exact interpolation. The maximum error is $> 10^7$. This is mathematics, nothing to do with rounding errors.

The second half of the figure goes back to the 1970s. Here there is an $(n-50)$-dimensional space of degree n polynomials that exactly interpolate the data. It is an *underdetermined least-squares problem*. Once the SVD (singular value decomposition) and associated algorithms were developed, a method came into play for choosing among these solutions by SVD-related regularization: set the numerically negligible singular values to zero, then compute a pseudoinverse. The figure was generated by the Matlab backslash command, which approximates the pseudoinverse, with $[-1, 1]$ discretized by 1000 points and a Chebyshev basis employed for numerical stability.

Did anyone ever put together the two halves of Figure 1 before 2019?

What we see here is the phenomenon of *double descent*, a term coined by Belkin, Hsu, Ma and Mandal that is having wide impact in machine learning. For $0 \le n \le 20$, the accuracy improves with n for a while, before increasing. Then, after the peak, it improves again for $50 \le n \le 58$: the second descent phase. In the large space of mathematically valid solutions, regularization is making a choice that tames Runge's oscillations.

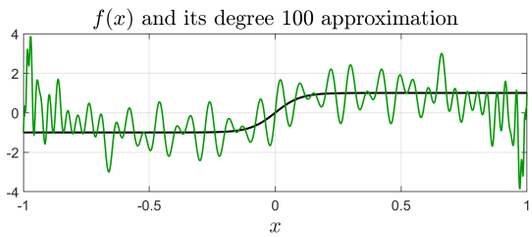

Figure 2. With $n \gg 50$, the huge errors are tamed.

The error doesn't diminish to 0 as $n \to \infty$. Figure 2 shows that the fit actually isn't very close, though it gets better if you penalize the higher polynomial degrees with a factor like $(1 + n)^{-1}$ (not shown). For a simple one-dimensional approximation problem like this, underdetermined least-squares cannot compete with more targeted methods. But in the unimaginably larger and more complicated multidimensional world of deep learning, the idea of going deep into the underdetermined, over-parametrized regime is proving exciting and powerful. As Belkin says: sometimes one can "fit without fear".

FURTHER READING

[1] M. Belkin, Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation, *Acta Numer.* (2021), 203–248.

Nick Trefethen
Trefethen is Professor of Applied Mathematics in Residence at Harvard University.