Double descent in the condition number

Tomaso Poggio[1], **Gil Kur, Andy Banburski**[1]

[1]Center for Brains, Minds, and Machines, MIT

This manuscript was compiled on April 26, 2020

In solving a system of n linear equations in d variables $Ax = b$, the condition number of the n, d matrix A measures how much errors in the data b affect the solution x. Estimates of this type are important in many inverse problems. An example is machine learning where the key task is to estimate an underlying function from a set of measurements at random points in a high dimensional space and where low sensitivity to error in the data is a requirement for good predictive performance. Here we discuss the simple observation, which is known but surprisingly little quoted (see Theorem 4.2 in (1)): when the columns of A are random vectors, the condition number of A is highest if $d = n$, that is when the inverse of A exists. An overdetermined system ($n > d$) as well as an underdetermined system ($n < d$), for which the pseudoinverse must be used instead of the inverse, typically have significantly better, that is lower, condition numbers. Thus the condition number of A plotted as function of d shows a double descent behavior with a peak at $d = n$.

Learning Theory | Double descent | Deep networks | Regression | Classification

The concept of condition number was introduced by Turing in 1948 (2) and has since played a key role in the theory of algorithms. The condition number of a function measures how much the output value of the function can change for a small change in the input argument. The condition number most commonly associated with $Ax = b$ is defined as the ratio of the relative error in x to the relative error in the data b. In terms of the l_2 norm on x and b, this leads to the following definition for the the condition number of A, denoted by $\kappa(A) = ||A||||A^\dagger||$ with $||A||$ being the operator norm of the m, n matrix A and $A^\dagger$ the pseudoinverse. The operator norm is defined as $||A|| = \sup_x ||Ax||$ with $||x|| = 1$. it is easy to see that $\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$ is the ratio of the maximal and minimal singular values of A.

The plot in the Figure 1 can be easily checked by calling the function "cond" in MatLab. The double descent pattern is apparently quite robust to choices of d and n, such that their ratio $\gamma = \frac{n}{d}$ is the same. The fact that the worse conditioning occurs when the inverse exists uniquely ($\gamma = 1$) seems at first surprising. This simple observation must have been realized by many. The proof is also simple because of a well-known characterization of the eigenvalues of random matrices (3). In fact, consider a $n \times d$ random matrix A.

We characterize its condition number by using the Marchenko–Pastur semi-circle law, which describes the asymptotic behavior of singular values of large rectangular random matrices. We assume that the entries of A are i.d.d. random variables with mean zero and variance one. We consider the limit for $n \to \infty$ with $\frac{n}{d} \to \gamma$.

Marchenko–Pastur claims that for $\gamma < 1$ the smallest and the largest singular values of $\frac{1}{d}AA^T$ are, respectively $(1 - \sqrt{\gamma})^2$ and $(1 + \sqrt{\gamma})^2$. For $\gamma > 1$ the largest and the smallest eigenvalues of $\frac{1}{n}A^TA$ are $(1 + \sqrt{\gamma^{-1}})^2$ and $(1 - \sqrt{\gamma^{-1}})^2$.

When $\gamma = 1$, and the entries are i.i.d. sub-Gaussian, the maximal singular value is concentrated around 2, but the minimal one is $\min\{n^{-1}, d^{-1}\}(\max\{\sqrt{n} - \sqrt{d-1}, \sqrt{d} - \sqrt{n-1}\})^2$ was observed in (4), for normal random variables it was first observed in (5).

Fo a system of linear equations $Ax = b$, when $n \approx d$, it is better to reduce/increase the data/variables (i.e. "better" to have more variables than data). The condition number associated with the minimum norm solution $x = A^\dagger b$ is usually much better – that is closer to 1 – than the condition number of a well-determined system with $n = d$, if the matrix A is random (see for instance (6, 7)).

There are interesting observations for machine learning. The most obvious is that kernel methods, which are a popular workhorse in machine learning, *do not require regularization in order to be well-conditioned*, if the kernel matrices are based on high dimensional i.i.d data, especially when $\gamma < 1$. This claim follows from recent results on kernels. The simplest form of the kernel matrix $K(x_j, x_i)$ is $K = XX^T$. We consider random matrices whose entries are $K(x_i^T x_j)$ with i.i.d. vectors x_i in $\mathbf{R}^p$ with normalized distribution (in Figure 2 we consider a radial kernel $K(||x_i - x_j||^2)$ for which similar arguments are likely to hold). Assuming that f is sufficiently smooth and the distribution of x_i's is sufficiently nice, El Karoui (8) showed that the spectral distributions of kernel dot-product matrices $K(x_i, x_j) = f(\frac{1}{d}XX^T)$ behave as if f is linear in the Marchenko–Pastur limit. In fact, El Karoui showed that under mild conditions, the kernel matrix is asymptotically equivalent to a linear combination of XX^T, the all-1's matrix, and the identity, and hence the limiting spectrum is Marcenko-Pastur. As a consequence, the claims about the condition number of a random matrix A also apply to kernel matrices with random data, see Figure 2.

More intriguing is the fact that the behavior of the condition number of $K^\dagger$ is similar to the *double descent* behavior of the test error by linear and kernel interpolants, which after pioneering work by Belkin ((9), see also (10)) has recently attracted much attention (9, 11–16). We will address the key role of stability for the theory of machine learning in a separate paper.

1. Burgisser P, Cucker F (2013) *Condition: The Geometry of Numerical Algorithms.* (Springer Publishing Company, Incorporated).
2. Turing AM (1948) Rounding-off errors in matrix processes. *Quarterly J. Mech. Appl. Math.* 1:287–308.
3. Marchenko VA, Pastur LA (1967) Distribution of eigenvalues for some sets of random matrices. *Mat. Sb. (N.S.)* 72(114):4:457–483.
4. Rudelson M, Vershynin R (2009) Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 62(12):1707–1739.
5. Davidson KR, Szarek SJ (2001) Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces* 1(317-366):131.
6. Chen Z, Dongarra JJ (2005) Condition numbers of gaussian random matrices. *SIAM J. Matrix Analysis Applications* 27:603–620.

All developed the idea.

The authors declare that they have no competing interests.

[2]To whom correspondence should be addressed. E-mail: tp@ai.mit.edu

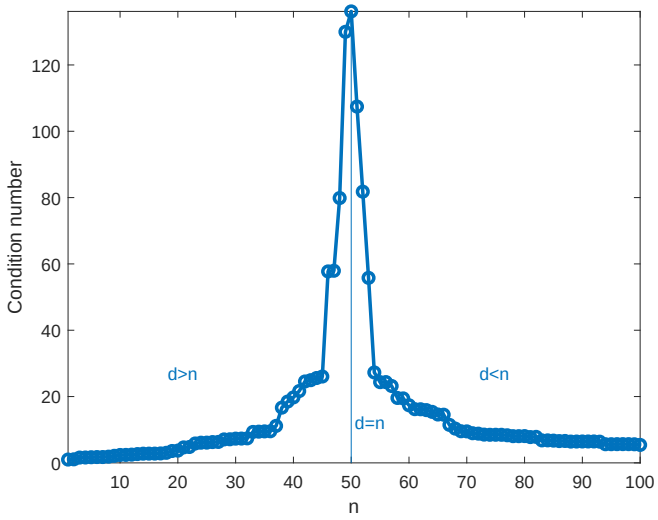

Fig. 1. The typical "double descent" of a $n \times d$ matrix with $\mathcal{N}(0,1)$ independent entries. The condition number is the worst when $n = d$.

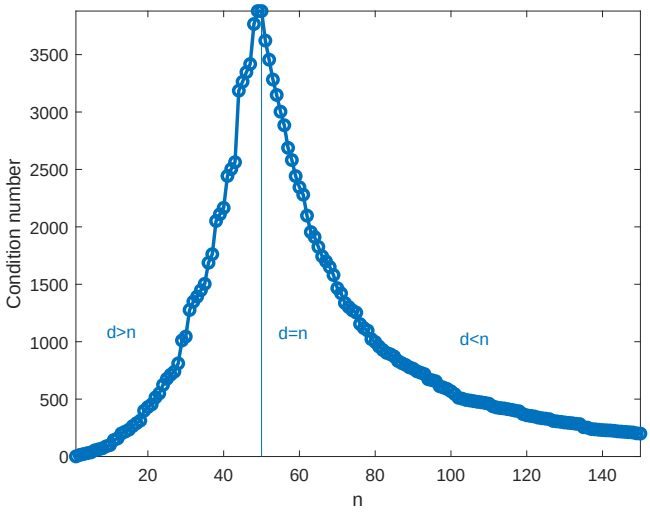

Fig. 2. Typical "double descent" of the condition number of the matrix $K(x_i, x_j)$, where $K(x, x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$ (radial kernel) and $x_1, \dots x_n$ are i.i.d. $\mathcal{N}(0, I_{d\times d})$. The condition number exhibts the same behavior as in the linear case (here $\sigma = 5$).

7. Hastie T, Montanari A, Rosset S, Tibshirani RJ (2019) Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.
8. El Karoui N (2010) The spectrum of kernel random matrices. *arXiv e-prints* p. arXiv:1001.0492.
9. Belkin M, Hsu D, Ma S, Mandal S (2019) Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* 116(32):15849–15854.
10. Advani MS, Saxe AM (2017) High-dimensional dynamics of generalization error in neural networks. *arXiv e-prints* p. arXiv:1710.03667.
11. Belkin M, Hsu D, Xu J (2019) Two models of double descent for weak features. *CoRR* abs/1903.07571.
12. Belkin M, Ma S, Mandal S (2018) To understand deep learning we need to understand kernel learning. *ArXiv e-prints*.
13. Mei S, Montanari A (2019) The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv e-prints* p. arXiv:1908.05355.
14. Rakhlin A, Zhai X (2018) Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon. *arXiv e-prints* p. arXiv:1812.11167.
15. Liang T, Rakhlin A (2018) Just Interpolate: Kernel "Ridgeless" Regression Can Generalize. *arXiv e-prints* p. arXiv:1808.00387.
16. Hastie T, Montanari A, Rosset S, Tibshirani RJ (2019) Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *arXiv e-prints* p. arXiv:1903.08560.