CENTER FOR
Brains
Minds+
Machines

**CBMM Memo No. 102** **February 5, 2020**

# Double descent in the condition number

**Tomaso Poggio**[1], **Gil Kur**[1], **Andy Banburski**[1]
[1]Center for Brains, Minds, and Machines, MIT

## Abstract

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 $x$. Bounds 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 well-known but surprisingly little quoted, that 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 $d = n$.

**This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.**

# Double descent in the condition number

Tomaso Poggio[1], Gil Kur, Andy Banburski

February 8, 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 $x$. Bounds 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 well-known but surprisingly little quoted (see Theorem 4.2 in [1]) that 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 $d = n$.**

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 the $n, d$ random matrix $A$. We characterize its condition number by using the Marchenko–Pastur law, which describes the asymptotic behavior of singular values of large rectangular random matrices. We assume that the entries of $A$ are independent, identically distributed random variables with mean 0 and variance $\sigma^2$. 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 $\min\{n^{-1}, d^{-1}\}(\max\{\sqrt{n}-\sqrt{d-1}, \sqrt{d}-\sqrt{n-1}\})^2$, as was observed in [4].

For the system of linear equations $Ax = b$, the implication is that is 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 [5]).

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 $\mathbb{R}^p$ 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 [6] 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 ([7], see also [8]) has recently attracted much attention [9, 10, 7, 11, 12, 13, 14]. We will address the key role of stability, discussed in some detail in previous versions of this memo, in a separate paper currently in preparation.

**Acknowledgments** We thank Gill Strang for very useful comments and for encouraging to publish this note. We are grateful to Felipe Cucker, Misha Belkin, lorenzo Rosasco, Aleksander Madry and especially Silvia Valle. This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216, and part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. This research was also sponsored by grants from the National Science Foundation (NSF-0640097, NSF-0827427), and AFSOR-THRL (FA8650-05-C-7262).

**Competing Interests** The authors declare that they have no competing financial interests.

**Correspondence** Correspondence and requests for materials should be addressed to T.Poggio (email: tp@ai.mit.edu).

**Author Contribution** All developed the basic theory.

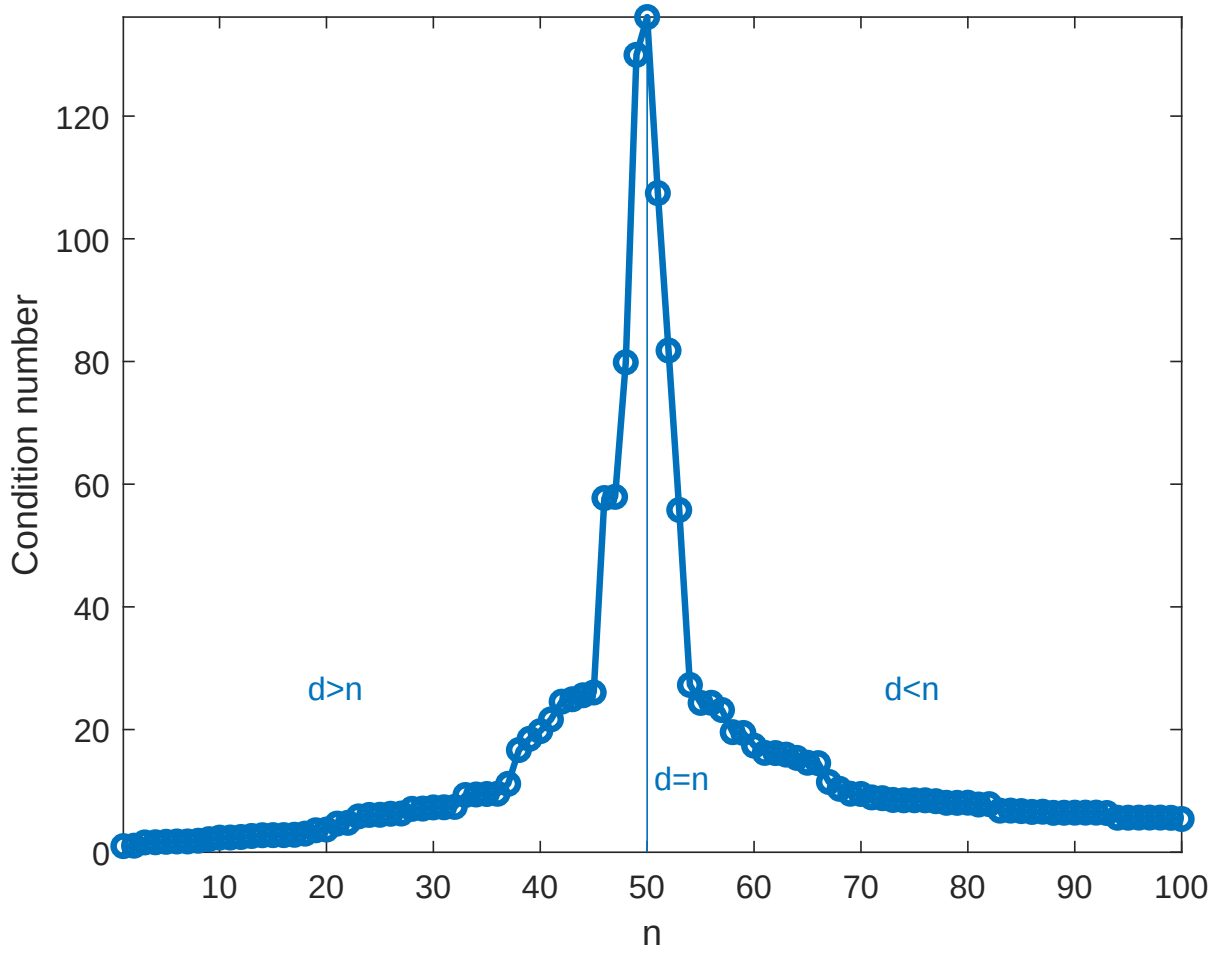


Figure 1: Typical double descent of the condition number (y axis) of a random data matrix distributed as $\mathcal{N}(0,1)$: the condition number is worse when $n = d$, better if $n > d$ (on the right of $n = d$) and also better if $n < d$ (on the left of $n = d$).

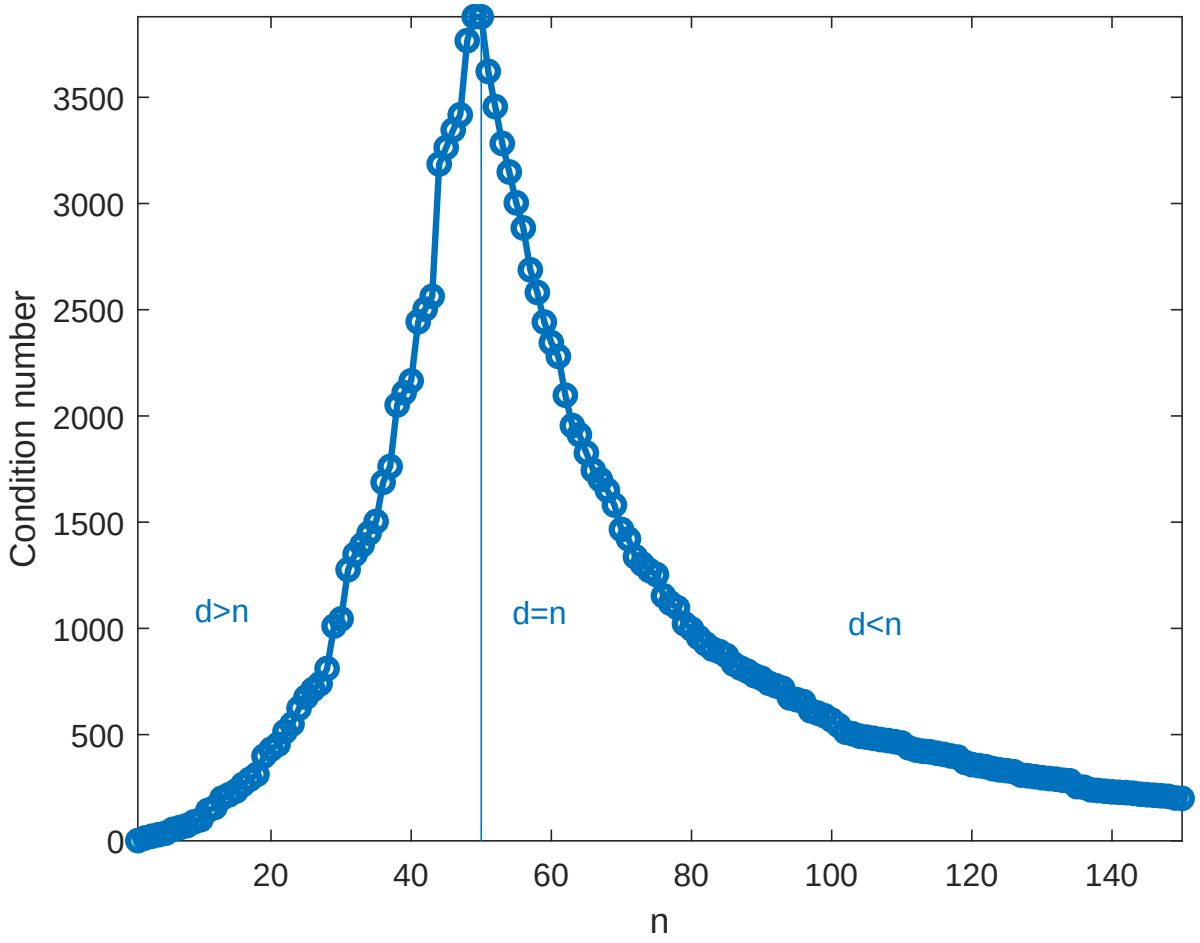


Figure 2: Typical double descent of the condition number (y axis) of a radial basis function kernel $K(x, x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$ built from a random data matrix distributed as $\mathcal{N}(0, 1)$: as in the linear case, the condition number is worse when $n = d$, better if $n > d$ (on the right of $n = d$) and also better if $n < d$ (on the left of $n = d$). The parameter $\sigma$ was chosen to be 5.

## References


[1] Peter Burgisser and Felipe Cucker. *Condition: The Geometry of Numerical Algorithms*. Springer Publishing Company, Incorporated, 2013.

[2] A. M. Turing. Rounding-off errors in matrix processes. *Quarterly J. Mech. Appl. Math.*, 1:287–308, 1948.

[3] Distribution of eigenvalues for some sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):4:457–483, 1967.

[4] Mark Rudelson and Roman Vershynin. 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, 2009.

[5] Zizhong Chen and Jack J. Dongarra. Condition numbers of gaussian random matrices. *SIAM J. Matrix Analysis Applications*, 27:603–620, 2005.

[6] Noureddine El Karoui. The spectrum of kernel random matrices. *arXiv e-prints*, page arXiv:1001.0492, Jan 2010.

[7] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[8] Madhu S. Advani and Andrew M. Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv e-prints*, page arXiv:1710.03667, Oct 2017.

[9] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *CoRR*, abs/1903.07571, 2019.

[10] M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. *ArXiv e-prints*, Feb 2018.

[11] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv e-prints*, page arXiv:1908.05355, Aug 2019.

[12] Alexander Rakhlin and Xiyu Zhai. Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon. *arXiv e-prints*, page arXiv:1812.11167, Dec 2018.

[13] Tengyuan Liang and Alexander Rakhlin. Just Interpolate: Kernel "Ridgeless" Regression Can Generalize. *arXiv e-prints*, page arXiv:1808.00387, Aug 2018.

[14] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *arXiv e-prints*, page arXiv:1903.08560, Mar 2019.