CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 102 **December 11, 2019**

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 report the simple observation that when the columns of A are random vectors, the condition number of A is highest, that is worse, if $d = n$, that is when the inverse of A exists. An overdetermined system ($n > d$) and especially 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 $d = n$. We discuss implications to recent machine learning developments.

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], Gill Kur, Andy Banburski

December 31, 2019

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 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 (see Box1) for the the condition number of A, denoted by $\kappa(A = ||A||||A^\dagger||$, where $||A||$ is the operator norm of the m, n matrix A is defined in terms of the vector norm of K^n, K^m as $||A|| = \sup_{x \in K^n, x \neq 0} \frac{||Ax||}{||x||}$ and $A^\dagger$ is the pseudoinverse. It is easy to see that $\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$ that 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 σ^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 $\gamma = 2$, 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 (see Box1), 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 3 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 $K(x_i, x_j) = f(XX^T)$ 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 3.

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 initial work by Belkin ([7], see also [8]) has recently attracted much attention [9, 10, 7, 11, 12, 13, 14]. It is natural to expect that some measure of stability of the interpolant solution should play a key role in determining the prediction error. We know that in the "classical" regime of fixed hypothesis space and $n \to \infty$, stability, defined as *cross-validation leave-one-out* (CV_{loo}) error, is important. We expect a similar notion of stability to be required in the "modern" high dimensional regime of $\frac{n}{d} \to \infty$, in which the minimum norm pseudoinverse plays a key role. In both cases, well-posedness, that is existence, uniqueness and especially stability of the solution, are the key requirement for predictivity. Stability is usually guaranteed during minimization of the empirical loss by complexity control under the form of (possibly) vanishing regularization (as in the definition of the pseudoinverse) or as implicitely provided by iterative gradient descent [15]. The notion of CV_{loo} stability (defined as the difference between the error made by the predictor obtained by using ERM on the training set S vs. the error of the predictor obtained from a slightly perturbed training set S^i) turns out to be necessary and sufficient for distribution independent generalization and consistency in the classical framework of ERM with a fixed hypothesis space [16, 17]. Here we can use a special case of CV_{loo} stability called uniform stability

[18] and defined as

Definition 1 *An algorithm has uniform stability β with respect to the loss function V if* $\forall S \in Z^n$, $\forall i \in \{1, \cdots, n\}, \quad \sup_{z \in Z^n} |V(f_S, z) - V(f_S^i, z)| \leq \beta$.

We derive a straightforward bound for interpolating kernel regressors for appropriate distributions of the data as (assuming Lipschitz properties of the loss function and $k(x, x) = \mathbf{k}^2$)

$$\beta \leq ||\frac{C}{n} k(x, X) K^{-1} Y|| \leq \frac{C}{n} ||k(x, X)|| ||K^{-1}|| ||Y|| = C k(x, x) \sigma_{min}^{-1} Y \leq \mathbf{k}^2 C \sigma_{min}^{-1} \tag{1}$$

where $\mathbf{k}$ and C do not depend on n, d of the training set S but σ_{min}^{-1} does (when $\frac{n}{d} < 1$ then $\sigma_{min} = (1 - \sqrt{\frac{n}{d}})^2$) . Thus $\beta \leq \mathbf{k}^2 C \sigma_{min}^{-1}$. Using Equation 1.3 in [19], we have with probability $1 - \delta$ that the expected error of the algorithm is

$$R(A) \leq (\sqrt{\beta B} + \frac{B}{\sqrt{(n)}}) \sqrt{\log(\frac{1}{\delta})} \tag{2}$$

where B is a uniform bound on the loss function. Thus we obtain

$$R(A) \leq (\sqrt{B \mathbf{k}^2 C (1 - \sqrt{\frac{n}{d}})^{-2}} + \frac{B}{\sqrt{n}}) \sqrt{\log(\frac{1}{\delta})}. \tag{3}$$

It is interesting to observe that β *and* $R(A)$ *are controlled by the minimum norm solution for any given* n, d. This is true *in the modern and also in the classical regime*, as it is clear by rewriting the bound on the expected error for the general case of regularized kernel regression with a regularization parameter λ:

$$R(A) - R_{emp}(A) \leq (\sqrt{T \mathbf{k}^2 C ||(K + n\lambda I)^\dagger||} + \frac{T}{\sqrt{n}}) \sqrt{\log(\frac{1}{\delta})}. \tag{4}$$

For d fixed, λ has to go to zero slower than $\frac{1}{n}$ in order for the right-hand side to go to zero. In this regime $\beta \to 0$ with $n \to \infty$ in a distribution-independent way, implies uniform convergence and compactness of the hypothesis space: the interpolation error is not zero but the generalization gap goes to zero. For $\frac{n}{d}$ fixed and $n \to \infty$, λ can be set to $\lambda = 0$; the minimum expected error then follows from the minimum of β, which itself follows from the minimum norm property of the pseudoinverse $||K^\dagger||$. The bound shows the double-descent property with respect to n, d when $||K^\dagger||$ does, for instance for random data, and thus in a distribution-dependent way. In summary, the property of uniform (or CV_{loo}) *stability* provides the crucial link between (minimum) norm and expected error.

In the modern regime, the expected error $R(A)$ remains different from the zero empirical error (interpolation) also for $n \to \infty$. An example is shown in Figure 3, demonstrating that even this simple estimate seems to capture the "double-descent" behavior of the test error.

Of course these bounds are loose. The condition number of the kernel matrix is not sufficient by itself to accurately estimate the out-of-sample error. An elegant and complex estimate of the

test error for kernel interpolators has been recently given [20, 21] adding intriguing details to the basic double descent behavior described by our simple analysis.

The open question is whether a stability condition such as *uniform stability* can also provide the expected behavior for the case of overparametrized deep neural networks. It has been shown recently [22, 23, 24] that with the exponential loss, gradient descent induces a dynamics such that the weight matrix for each layer of the network converges to a minimum norm solution. In fact, for a K-layer network a simple bound on on β is given by $C||W_K|| \cdots ||W_1||$. As we mentioned, the weights $W_1 \cdots W_K$ to which gradient descent converges are the minimum norm solutions (for margin greater or equal to 1) for each layer. Then, following the same logic we used above for kernel interpolants, the expected error – assuming zero-error interpolation on the data – is $R(A) \leq (\sqrt{TC'||W_K|| \cdots ||W_1||} + \frac{T}{\sqrt{n}})\sqrt{\log(\frac{1}{\delta})}$ and is a minimum because each of the W_k is a minimum norm solution.

Acknowledgments We thank Gil Strang and Felipe Cucker and Lorenzo Rosasco 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.

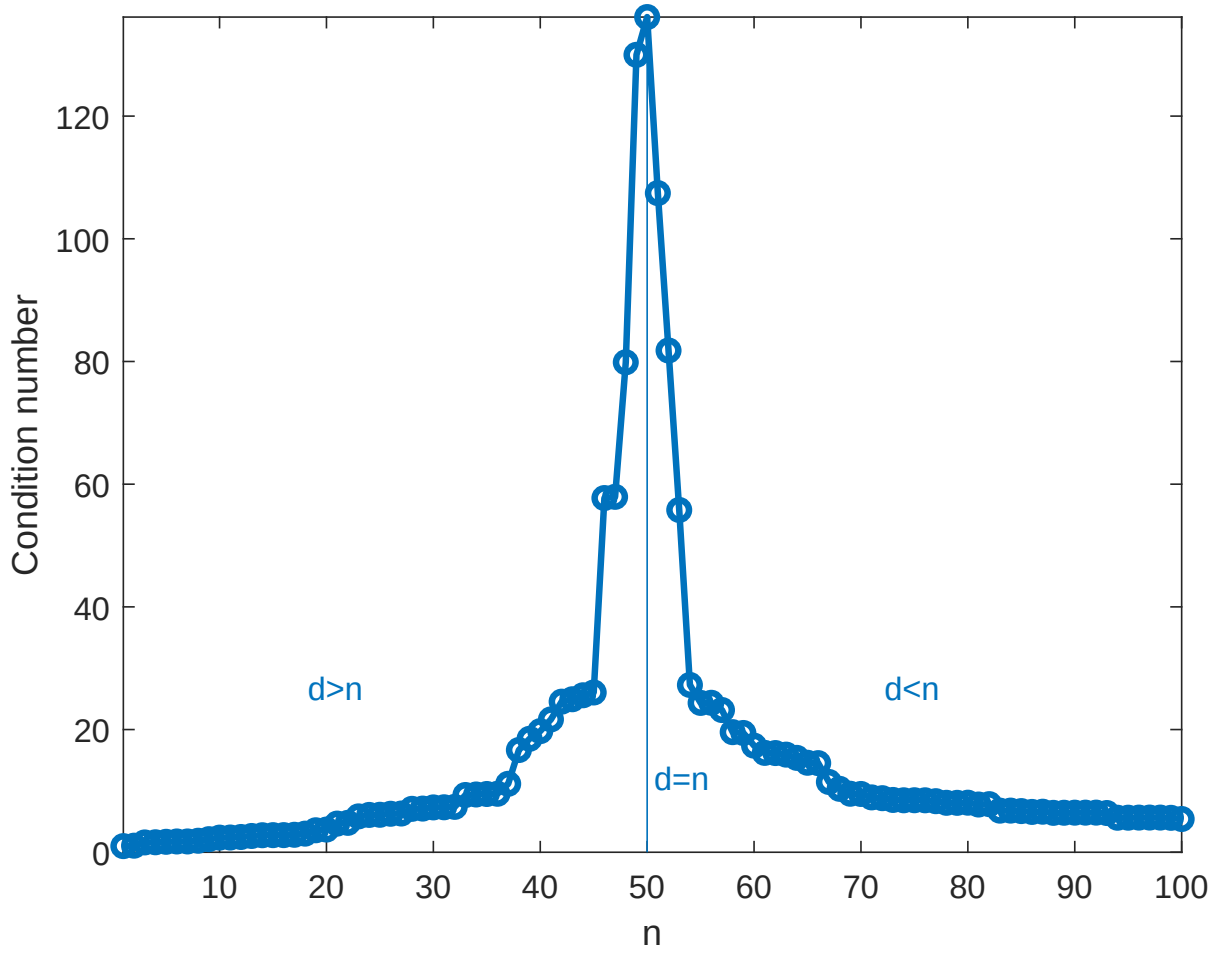

Figure 1: Typical double descent of the condition number (y axis) of a random data matrix distributed as $\mathcal{N}(0,1)$: the condition number is worse when $n = d$, better if $n > d$ (on the right of $n = d$) and also better if $n < d$ (on the left of $n = d$).

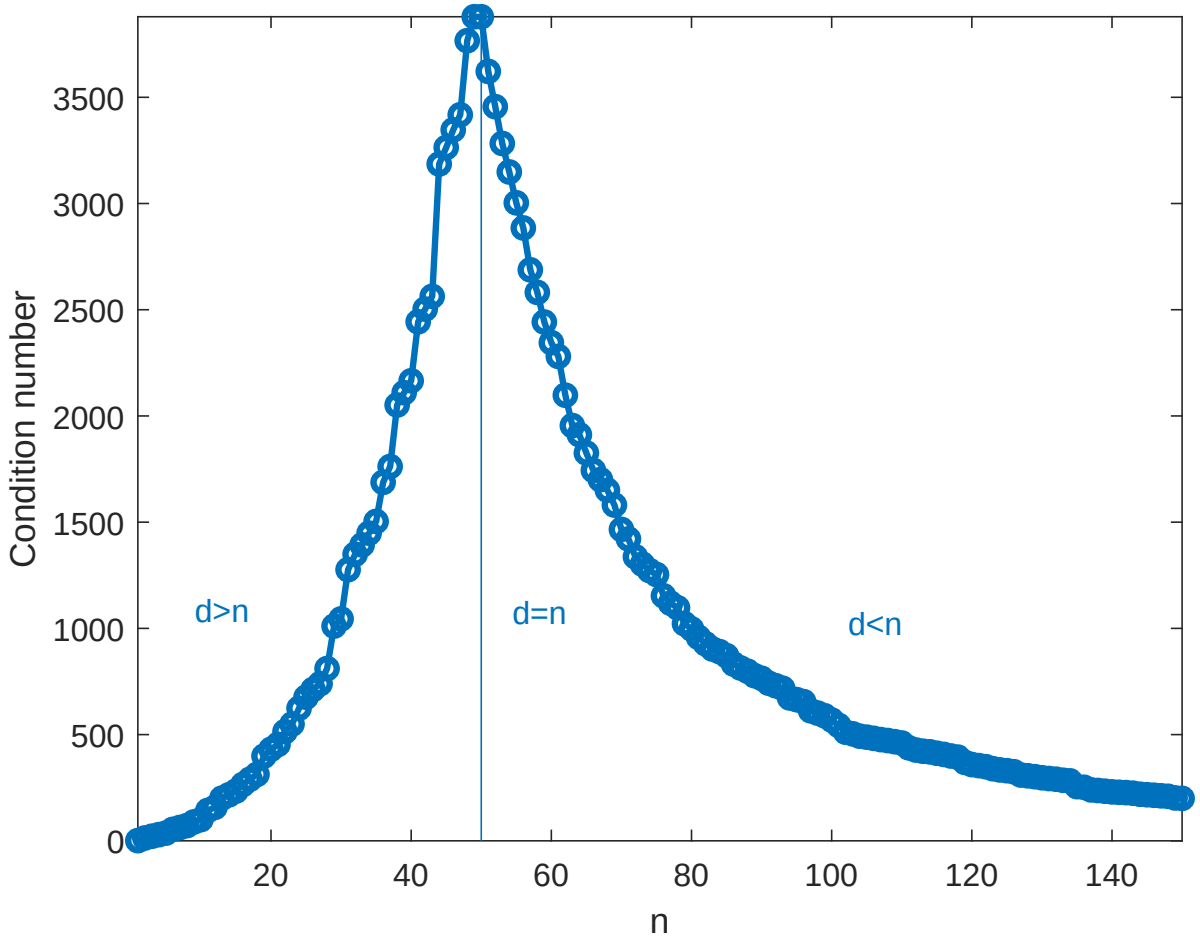

Figure 2: Typical double descent of the condition number (y axis) of a radial basis function kernel $K(x, x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$ built from a random data matrix distributed as $\mathcal{N}(0, 1)$: as in the linear case, the condition number is worse when $n = d$, better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

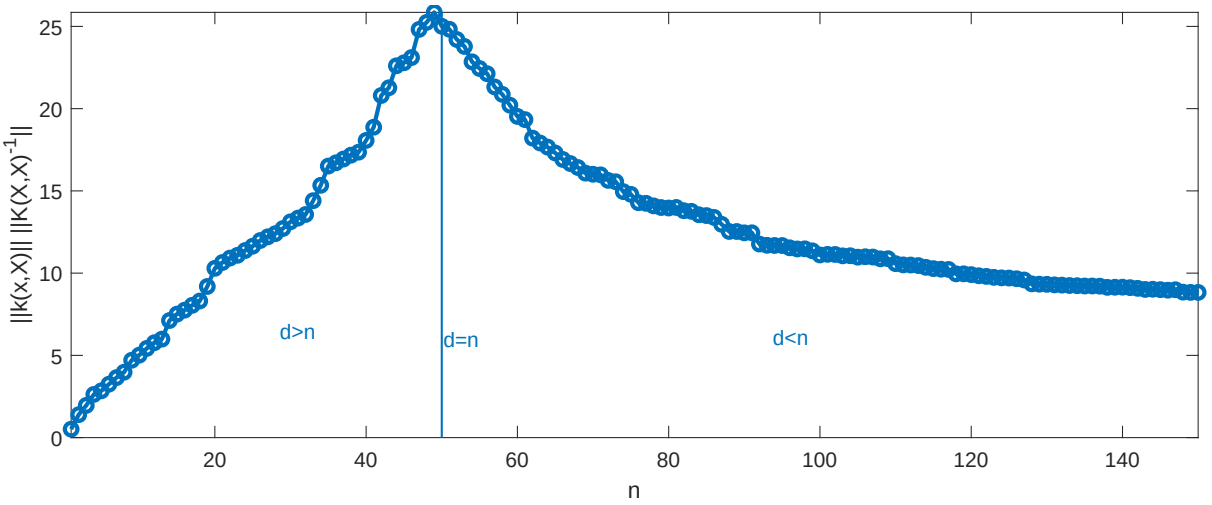

Figure 3: Plot of $||K(x,X)||\ ||K(X,X)^{\dagger}||$, where x is a out-of-sample data point (not in X) and K is a radial basis function kernel $K(x,x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$, plotted with $\sigma = 8$. Both the dataset X and the out-of-sample point were generated from the $\mathcal{N}(0,1)$ distribution.

Boxes

Box1: ERM and Kernel Machines

Consider "learning the function f from data $S = (x_1, y_1; x_2, y_2, \cdots x_n, y_n)$ by computing

$$\min_{f \in B_R} \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2. \tag{5}$$

We assume that $f(\mathbf{x}) = \sum_{1=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x})$ and that f is in the ball B_R of radius R in $\mathcal{H}$ (eg $\|f\|_K \leq R$). Then $\mathcal{H} = \overline{I_K(B_R)}$ is compact – where $I_K : \mathcal{H}_K \hookrightarrow C(X)$ is the inclusion and $C(X)$ is the space of continuous functions with the sup norm [25]. In this case the minimizer of the generalization error $I[f]$ is well-posed. Minimization of the empirical risk (Equation (5)) is also well-posed: it provides a set of linear equations to compute the coefficients $\mathbf{c}$ of the solution f as

$$K\mathbf{c} = \mathbf{y} \tag{6}$$

where $\mathbf{y} = (y_1, ..., y_n)$ and $(K)_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$. Notice that this last set of linear equations is well-posed even without the constraint $||f||_K \leq R$: if K is symmetric and positive definite and the x_i are distict the K^{-1} exists and $||f||_K^2$ is automatically bounded, with a bound that increase with n. For any fixed n the condition number of K is finite. A regularized form of ERM is

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 + \lambda \|f\|_K^2, \tag{7}$$

which gives the following set of equations for $\mathbf{c}$ (with $\lambda \geq 0$)

$$(K + n\lambda I)\mathbf{c} = \mathbf{y}, \tag{8}$$

which for $\lambda = 0$ reduces to Equation (6). In both cases, stability of the empirical risk minimizer provided by Equation (7) can be characterized using the classical notion of *condition number* of the problem. The change in the solution f due to a variation in the data $\mathbf{y}$ can be bounded as $\frac{\|\Delta f\|}{\|f\|} \leq \|K + n\lambda I\| \left\|(K + n\lambda I)^{-1}\right\| \frac{\|\Delta \mathbf{y}\|}{\|\mathbf{y}\|}$ where the condition number $\|K + n\lambda I\| \left\|(K + n\lambda I)^{-1}\right\|$ is controlled by $n\lambda$. A large value of $n\lambda$ gives condition numbers close to 1, whereas ill-conditioning may result if $\lambda = 0$ and the ratio of the largest to the smallest eigenvalue of K is large. Though this was the classical argument, it is now clear (because of recent results such as El Karoui [6]) that random K matrices are typically well-conditioned even for $\lambda = 0$. In other words, for i.i.d high-dimensional data,

$$\frac{\|\Delta f\|}{\|f\|} \leq \|K\| \left\|(K)^{\dagger}\right\| \frac{\|\Delta \mathbf{y}\|}{\|\mathbf{y}\|}, \tag{9}$$

and the condition number $\kappa(K) = \|K\| \left\|(K)^{\dagger}\right\|$ is close to 1, *especially if d is different from n.*

Box2: Classical Learning Theory

In the classical setting, a key property of a learning algorithm is *generalization*: the empirical error must converge to the expected error when the number of examples n increases to infinity, while the class of functions $\mathcal{H}$, called the *hypothesis space*, is kept fixed. An algorithm that guarantees good generalization will predict well, if its empirical error on the training set is small. Empirical risk minimization (ERM) on $\mathcal{H}$ represents perhaps the most natural class of learning algorithms: the algorithm selects a funcion $f \in \mathcal{H}$ that minimizes the empirical error – as measured on the training set.

One of the main achievements of the classical theory was a complete characterization of the necessary and sufficient conditions for generalization of ERM, and for its *consistency* (consistency requires asymptotic convergence of the expected risk to the minimum risk achievable by functions in $\mathcal{H}$; for ERM generalization is equivalent to consistency). It turns out that consistency of ERM is equivalent to a precise property of the hypothesis space: $\mathcal{H}$ has to be a *uniform Glivenko-Cantelli (uGC)* class of functions.

Later work showed that an apparently separate requirement – the well-posedness of ERM – is in fact equivalent to consistency of ERM. Well-posedness usually means *existence, uniqueness and stability* of the solution. The critical condition is stability of the solution. Stability is equivalent to some notion of continuity of the learning map (induced by ERM) that maps training sets into the space of solutions, eg $L : Z^n \to \mathcal{H}$. In particular, it was proved [16, 17] that *distribution-independent CV_{loo} stability guarantees generalization and for ERM is equivalent to consistency.*

We recall the definition of *leave-one-out cross-validation (in short, CV_{loo}) stability*:

$$\forall i \in \{1, \dots, n\} \;\; P_S \left\{ |V(f_S, z_i) - V(f_{S^i}, z_i)| \leq \beta_{CV} \right\} \geq 1 - \delta_{CV}, \tag{10}$$

where $V(f, z)$ is a loss function that is Lipschitz and bounded for the range of its arguments and $z = ((x, y)$. CV_{loo} stability measures the difference between the errors at a point z_i when it is in the training set S of f_S wrt when is not. The definition of CV_{loo} was introduced to deal with general situations in which $\mathcal{H}$ may not have a norm. The definition of leave-one-out stability is simpler in the framework of inverse problems when $\mathcal{H}$ is a RKHS and the noise in the data can be assumed to affect only the "outputs" y_i. Then a condition number can be defined: a good condition number close to 1 implies then good CV_{loo} stability. Both definitions capture the basic idea of stability of a well-posed problem: the function "learned" from a training set should, with high probability, change little in its pointwise predictions for a small change in the training set, such as deletion of one of the examples or label noise affecting some of the training data.

In the modern regime, in which both n and d grow to infinity, the generalization gap does not go to zero. The classical approach – of asymptotic generalization and then consistency – cannot be used because there is no fixed hypothesis space and the setup must be distribution-dependent. However, the requirement of well-posedness, including stability, should remain. The argument at the end of the main text may suggest how to formally show that uniform stability explains the main properties of both the classical and the modern regime including the distribution-dependent double descent property of the latter.

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.

[15] Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

[16] Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006.

[17] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, March 2004.

[18] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal Machine Learning Research*, 2001.

[19] Olivier Bousquet, Yegor Klochkov, and Nikita Zhivotovskiy. Sharper bounds for uniformly stable algorithms. *arXiv e-prints*, page arXiv:1910.07833, Oct 2019.

[20] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize, 2018.

[21] Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the Risk of Minimum-Norm Interpolants and Restricted Lower Isometry of Kernels. *arXiv e-prints*, page arXiv:1908.10292, Aug 2019.

[22] A. Banburski, Q. Liao, B. Miranda, T. Poggio, L. Rosasco, B. Liang, and J. Hidary. Theory of deep learning III: Dynamics and generalization in deep networks. *CBMM Memo No. 090*, 2019.

[23] Mor Shpigel Nacson, Suriya Gunasekar, Jason D. Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and Depth-Sensitive Margins in Homogeneous and Non-Homogeneous Deep Models. *arXiv e-prints*, page arXiv:1905.07325, May 2019.

[24] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *CoRR*, abs/1906.05890, 2019.

[25] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of AMS*, 39:1–49, 2001.