CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 108 **June 22, 2020**

For interpolating kernel machines, the minimum norm ERM solution is the most stable

Akshay Rangamani[1,2], **Lorenzo Rosasco**[1], **Tomaso Poggio**[1,2]

1: Center for Brains, Minds, and Machines
2: McGovern Institute for Brain Research at MIT, Cambridge, MA, USA

Abstract

We study the average CV_{loo} stability of kernel ridge-less regression and derive corresponding risk bounds. We show that the interpolating solution with minimum norm has the best CV_{loo} stability, which in turn is controlled by the condition number of the empirical kernel matrix. The latter can be characterized in the asymptotic regime where both the dimension and cardinality of the data go to infinity. Under the assumption of random kernel matrices, the corresponding test error follows a double descent curve.

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

For interpolating kernel machines, the minimum norm ERM solution is the most stable

Akshay Rangamani, Lorenzo Rosasco, Tomaso Poggio

June 22, 2020

Abstract

We study the average CV_{loo} stability of kernel ridge-less regression and derive corresponding risk bounds. We show that the interpolating solution with minimum norm has the best CV_{loo} stability, which in turn is controlled by the condition number of the empirical kernel matrix. The latter can be characterized in the asymptotic regime where both the dimension and cardinality of the data go to infinity. Under the assumption of random kernel matrices, the corresponding test error follows a double descent curve.

1 Introduction

Statistical learning theory studies the learning properties of machine learning algorithms, and more fundamentally under which conditions learning from finite data is possible. In this context, the classical theory focuses on the size of the hypothesis space in terms of different complexity measures, such as combinatorial dimensions, covering numbers and Rademacher/Gaussian complexities, see [1, 2] and references therein. Another more recent approach is based on defining suitable notions of stability with respect to perturbation of the data, see e.g. [3, 4]. In this view, the continuity of the process that maps data to estimators is crucial, rather than the complexity of the hypothesis space. Different notions of stability can be considered, depending on the data perturbation and metric considered, see [4] and references therein. Interestingly, the stability and complexity approaches to characterizing the *learnability* of problems are not at odds with each other, and can be shown to be equivalent as shown in [5] (see also [6]).

In modern machine learning, it is common to consider large and even possibly *infinite* models for which deriving sharp statistical learning results is challenging and has led to much work. This is the case with a number of learning problems, for instance kernel methods [7] corresponding to models with infinitely many parameters [8], for high dimensional learning with sparsity, where the number of parameters is much larger than the number of points, and especially for deep networks [9], where billions of parameters are common. In particular, studying the properties of deep networks led to the observation that learning is possible also when perfectly fitting/interpolating the data, a property often associated with overfitting and loss of learning accuracy [10]. This observation has led to much recent work trying to ground theoretically this empirical findings. As noted in [11], interpolation is not a property exclusive to deep neural networks, but is possible with other overparameterized models and in particular with kernel methods. These models are easier to study and their properties have been recently revisited, since classic results focus on situations where constraints or penalties are added, preventing interpolation. For example, high dimensional linear models are considered in [12, 13, 14], and unpenalized kernel least squares in [15, 16], which we also study in this paper.

Our main contribution in this paper, is to consider these questions through a stability approach. The stability properties of regularized kernel methods are well known, and indeed in this case strong guarantees can be established using the property of uniform stability [3]. These bounds, however, cannot be used in the limit of vanishing regularization. In unregularized problems, we consider here the minimal norm solution among

all those interpolating the data. It is well known that, the numerical stability of this solution is governed by the condition number of the associated kernel matrix (see discussion of why overparametrization is "good" in [17]). Our results shows that the condition number also controls stability in a statistical sense. Indeed, our main result shows that the the average stability of the minimum norm solution is controlled by the expectation of the kernel matrix. Using results from high dimensional statistics and random matrix theory, such a condition number can be controlled in the limit where the data size and dimension both go to infinity [18, 19]. Further, stability can be shown to directly control the excess risk, hence the test error. In this view, among all interpolating solutions, the one with minimal norm has the best stability and hence the best test error. In particular, the same conclusion is also true for gradient descent, since the it converges to the minimal norm solution in the setting we consider, see e.g. [20] and references therein.
The rest of the paper is organized as follows. In Section 2, we introduce basic ideas in statistical learning with empirical risk minimization. In Section 3, we recall some basic stability results, and finally in Section 4, we study the stability of minimum norm interpolating solutions.

2 Statistical Learning and Empirical Risk Minimization

We begin recalling the basic ideas in statistical learning theory. In this setting, there is an unknown probability distribution μ on the product space $Z = X \times Y$. In the following, we consider $X = \mathbb{R}^d$ and $Y = \mathbb{R}$. The distribution μ is fixed but unknown, and we are given a training set S consisting of n samples (thus $|S| = n$) drawn i.i.d. from the probability distribution on Z^n, $S = (z_i)_{i=1}^n = (\mathbf{x}_i, y_i)_{i=1}^n$. Intuitively, the goal of supervised learning is to use the training set S to "learn" a function f_S that evaluated at a new value $\mathbf{x}_{new}$ should predict the associated value of y_{new}, i.e. $y_{new} \approx f_S(\mathbf{x}_{new})$. If y is real-valued, we have regression. If $y \in \{-1, 1\}$, we have binary classification.
To define the problem more precisely, we measure goodness of a function, introducing a loss function V. We denote by $V(f, z)$ the price we pay when the prediction for a given $\mathbf{x}$ is $f(\mathbf{x})$ and the true value is y. Hence, the loss is a function $V : \mathcal{F} \times Z \to [0, \infty)$, where $\mathcal{F}$ is the space of measurable functions from X to Y. We also introduce a hypothesis space $\mathcal{H} \subseteq \mathcal{F}$ where the considered algorithms will search for solutions. With the above notation, the *expected error* of f is defined as,

$$I[f] = \mathbb{E}_z V(f, z)$$

which is the expected loss on a new sample drawn according to the data distribution. In this setting, statistical learning can be seen as the problem of finding an approximate solution of the problem

$$\min_{f \in \mathcal{H}} I[f] \tag{1}$$

given a training set S. A natural and classical approach to derive an approximate solution is empirical risk minimization (ERM). This approach is based on the simple idea of replacing the expected risk in (1) with the empirical risk hence deriving the problem

$$\min_{f \in \mathcal{H}} I_S[f] = \frac{1}{n} \sum_{i=1}^{n} V(f, z_i) \tag{2}$$

A natural error measure for our ERM solution f_S is the expected excess risk $\mathbb{E}_S[I[f_S] - \min_{f \in \mathcal{H}} I[f]]$. Another common error measure is the expected generalization error/gap given by $\mathbb{E}_S[I[f_S] - I_S[f_S]]$. These two error measures are closely related since, the expected risk is easily bounded by the expected generalization error (see Lemma 5).

2.1 Kernel Least Squares and Minimal Norm Solution

In this paper, we assume that the loss function V is the square loss, that is, $V(f,z) = (y - f(\mathbf{x}))^2$. Choosing the square loss ERM becomes,

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2. \tag{3}$$

The focus in this paper is on reproducing kernel Hilbert spaces, defined by a positive definite kernel $K : X \times X \to \mathbb{R}$ or an associated feature map $\Phi : X \to \mathcal{H}$, such that $K(\mathbf{x},\mathbf{x}') = \langle\Phi(\mathbf{x}),\Phi(\mathbf{x}')\rangle_{\mathcal{H}}$ for all $\mathbf{x},\mathbf{x}' \in X$, where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$. In this setting, functions are linearly parameterized, that is there exists $w \in \mathcal{H}$ such that $f(\mathbf{x}) = \langle w,\Phi(\mathbf{x})\rangle_{\mathcal{H}}$ for all $x \in X$. A simple yet relevant example are linear functions $f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}$, that correspond to $\mathcal{H} = \mathbb{R}^d$ and Φ the identity map. Problem 3 typically has multiple solutions, and the minimal norm solution that is

$$f_S^\dagger = \min_{f\in\mathcal{M}} \|f\|_{\mathcal{H}}\,, \qquad \mathcal{M} = \arg\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2. \tag{4}$$

Here $\|\cdot\|_{\mathcal{H}}$ is the norm on $\mathcal{H}$ induced by the inner product. The minimal norm solution can be shown to be unique and satisfy a representer theorem, that is for all $\mathbf{x} \in X$

$$f_S^\dagger(\mathbf{x}) = \sum_{i=1}^{n} K(\mathbf{x},\mathbf{x}_i)c_{S,i}, \qquad \mathbf{c}_S = \mathbf{K}^\dagger\mathbf{y} \tag{5}$$

where $\mathbf{c_S} = (c_{S,1},\ldots,c_{S,n}), \mathbf{y} = (y_1\ldots y_n) \in \mathbb{R}^n$, $\mathbf{K}$ is the n by n matrix with entries $\mathbf{K}_{ij} = K(\mathbf{x}_i,\mathbf{x}_j)$, $i,j = 1,\ldots,n$, and $\mathbf{K}^\dagger$ is the Moore-Penrose pseudoinverse of $\mathbf{K}$. If we assume $n \le d$ and to have n linearly independent input points, that is the rank of $\mathbf{X}$ is n, then it is possible to show that, for many kernels one can replace $\mathbf{K}^\dagger$ by $\mathbf{K}^{-1}$, see Remark 2. Note, invertibility is necessary and sufficient for interpolation $f_S^\dagger(\mathbf{x}_i) = y_i$ for all $i = 1,\ldots,n$, in which case error training error in (4) is zero. We illustrate this in the case of linear kernels/functions.

Remark 1 (Pseudoinverse for underdetermined linear systems) *For linear functions, problem* (3) *is simply the linear least squares method*

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top\mathbf{x}_i - y_i)^2. \tag{6}$$

If the rank of $\mathbf{X} \in \mathbb{R}^{d\times n}$ is n, then $\mathbf{w}^\top\mathbf{x}_i = y_i$ for all $i = 1,\ldots,n$, and the minimal norm solution, also called Moore-Penrose solution, is given by

$$(\mathbf{w}_S^\dagger)^\top = \mathbf{y}^\top\mathbf{X}^\dagger$$

where the pseudoinverse $\mathbf{X}^\dagger$ takes the form $\mathbf{X}^\dagger = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}$.

Remark 2 (Invertibility of translation invariant kernels) *Translation invariant kernels are a family of kernel functions given by $K(\mathbf{x}_1,\mathbf{x}_2) = k(\mathbf{x}_1 - \mathbf{x}_2)$ where k an even function on $\mathbb{R}^d$. Translation invariant kernels are Mercer kernels (positive semidefinite) if the Fourier transform of $k(\cdot)$ is non-negative. For Radial Basis Function kernels ($K(\mathbf{x}_1,\mathbf{x}_2) = k(||\mathbf{x}_1 - \mathbf{x}_2||)$) we have the additional property due to Theorem 2.3 of [21] that for distinct points $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n \in \mathbb{R}^d$ the kernel matrix $\mathbf{K}$ is non-singular and thus invertible.*

The above discussion is directly related to to regularization approaches.

Remark 3 (Stability and Tikhonov regularization) *Tikhonov regularization is used to prevent potential unstable behaviors. In the above setting, it corresponds to replacing Problem* (4) *by*

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

where the corresponding unique solution f_S^λ is given by

$$f_S^\lambda(\mathbf{x}) = \sum_{i=1}^{n} K(\mathbf{x}, \mathbf{x}_i)c_i, \qquad \mathbf{c} = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{y}.$$

In contrast to minimal norm solutions, the above approach prevents interpolation. The properties of the corresponding estimator are well known. In this paper, we complement these results focusing on the case $\lambda \to 0$.

Finally, we end recalling the connection between minimal norm and the gradient descent.

Remark 4 (Minimum norm and gradient descent) *In our setting, it is well know the both batch and stochastic gradient iterations converge exactly to the minimal norm solution, when multiple solutions exist, see e.g. [20]. Thus, a study of the properties of minimal norm solutions explain the properties of the solution to which gradient descent converges. In particular, when ERM has multiple interpolating solutions, gradient descent converges to the most stable one, as we show next.*

3 Error Bounds via Stability

In this section, we recall basic results relating the learning and stability properties of ERM. Throughout the paper, we assume that ERM achieves a minimum, albeit the extension to almost minimizer is possible [22] and important for exponential-type loss functions [23]. We do not assume the expected risk to achieve a minimum. Since we will be considering leave-one-out stability in this section, we look at solutions to ERM (2) over the complete training set $S = \{z_1, z_2, \ldots, z_n\}$ and the leave one out training set $S_i = \{z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$ The excess risk of ERM can be easily related to its stability properties. Here, we follow [22], and say that an algorithm is Cross-Validation leave-one-out (CV_{loo}) stable in expectation, if there exists $\beta_{CV} > 0$ such that for all $i = 1, \ldots, n$,

$$\mathbb{E}_S[V(f_{S_i}, z_i) - V(f_S, z_i)] \leq \beta_{CV}. \tag{7}$$

This definition is justified by the following result.

Lemma 5 (Excess Risk & CV_{loo} Stability) *For all $i = 1, \ldots, n$,*

$$\mathbb{E}_S[I[f_{S_i}] - \inf_{f \in \mathcal{H}} I[f]] \leq \mathbb{E}_S[V(f_{S_i}, z_i) - V(f_S, z_i)]. \tag{8}$$

We present the proof of the above Lemma in Appendix A.2 due to lack of space. Below we discuss some more aspects of stability and its connection to other quantities in statistical learning theory.

Remark 6 (CV_{loo} stability in expectation and in probability) *In [22], CV_{loo} stability is defined in probability, that is there exists $\beta_{CV}^P > 0$, $0 < \delta_{CV}^P \leq 1$ such that*

$$\mathbb{P}_S\{|V(f_{S_i}, z_i) - V(f_S, z_i)| \geq \beta_{CV}^P\} \leq \delta_{CV}^P.$$

Note that the absolute value is not needed for ERM since almost positivity holds [22], that is $V(f_{S_i}, z_i) - V(f_S, z_i) > 0$. Then CV_{loo} stability in probability and in expectation are clearly related and indeed equivalent for bounded loss functions. CV_{loo} stability in expectation (7) is what we study in the following sections.

Remark 7 (Connection to uniform stability and other notions of stability) *Uniform stability, introduced by [3], correspondsin our notation to the assumption that there exists $\beta_u > 0$ such that for all $i = 1, \ldots, n$, $\sup_{z \in Z} |V(f_{S_i}, z) - V(f_S, z)| \leq \beta_u$. Clearly this is a strong notion implying most other definitions of stability. We note that there are number of different notions of stability. We refer the interested reader to [4] , [22].*

Remark 8 (CV_{loo} Stability & Learnability) *A natural question is to which extent suitable notions of stability are not only sufficient but also necessary for controlling the excess risk of ERM. Classically, the latter is characterized in terms of uniform version of the law of large numbers, which itself can be characterized in terms of suitable complexity measures of the hypothesis class. Uniform stability is too strong to characterize consistency while CV_{loo} stability turns out to provide a suitably weak definition as shown in [22], see also [4], [22]. Indeed, a main result in [22] shows that CV_{loo} stability is equivalent to consistency of ERM:*

Theorem 9 [22] *For ERM and bounded loss functions, CV_{loo} stability in probability with β_{CV}^P converging to zero for $n \to \infty$ is equivalent to consistency and generalization of ERM.*

Remark 10 (CV_{loo} stability & in-sample/out-of-sample error) *Let $(S, z) = \{z_1, \ldots, z_n, z\}$, and the corresponding ERM solution $f_{(S,z)}$, then* (8) *can be equivalently written as,*

$$\mathbb{E}_S[I[f_S] - \inf_{f \in \mathcal{F}} I[f]] \leq \mathbb{E}_{S,z}[V(f_S, z) - V(f_{(S,z)}, z)].$$

Thus CV_{loo} stability measures how much the loss changes when we test on a point that is present in the training set and absent from it. In this view, it can be seen as an average measure of the difference between in-sample and out-of-sample error.

Remark 11 (CV_{loo} stability and generalization) *A common error measure is the (expected) generalization gap $\mathbb{E}_S[I[f_S] - I_S[f_S]]$. For non-ERM algorithms, CV_{loo} stability by itself not sufficient to control this term, and further conditions are needed* [22]*, since*

$$\mathbb{E}_S[I[f_S] - I_S[f_S]] = \mathbb{E}_S[I[f_S] - I_S[f_{S_i}]] + \mathbb{E}_S[I_S[f_{S_i}] - I_S[f_S]].$$

The second term becomes for all $i = 1, \ldots, n$,

$$\mathbb{E}_S[I_S[f_{S_i}] - I_S[f_S]] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_S[V(f_{S_i}, z_i) - V(f_S, z_i)] = \mathbb{E}_S[V(f_{S_i}, z_i) - V(f_S, z_i)]$$

and hence is controlled by CV stability. The first term is called expected leave one out error in [22] *and is controlled in ERM as $n \to \infty$, see Theorem 9 above.*

4 CV_{loo} Stability of Kernel Least Squares

In this section we will analyze the expected CV_{loo} stability 7 of interpolating in kernel least squares (3). We also compare the stabilty of the minimal norm interpolating solution (4) to the stabilty of the other interpolating solutions to the Kernel least squares problem.

Theorem 12 (Main Theorem) *Consider the kernel least squares problem* (3)*, with a bounded kernel and bounded outputs y, that is there exist $\kappa, M > 0$ such that*

$$K(\mathbf{x}, \mathbf{x}') \leq \kappa, \qquad |y| \leq M, \tag{9}$$

almost surely. Then,

$$\mathbb{E}_S[V(f_{S_i}^\dagger, z_i) - V(f_S^\dagger, z_i)] \leq C_1 \beta_1 + C_2 \beta_2 \tag{10}$$

Where $\beta_1 = \mathbb{E}_S \left[||\mathbf{K}^{\frac{1}{2}}||_{op} ||\mathbf{K}^\dagger||_{op} \times cond(\mathbf{K}) \times ||\mathbf{y}|| \right]$ and, $\beta_2 = \mathbb{E}_S \left[||\mathbf{K}^{\frac{1}{2}}||_{op}^2 ||\mathbf{K}^\dagger||_{op}^2 \times (cond(\mathbf{K}))^2 \times ||\mathbf{y}||^2 \right]$, and C_1, C_2 are absolute constants that do not depend on either d or n. In particular, the minimum norm interpolating solution(5)*, is also the most stable solution in the expected CV_{loo} sense.*

In the above theorem $||\mathbf{K}||_{op}$ refers to the operator norm of the kernel matrix $\mathbf{K}$, $||\mathbf{y}||$ refers to the standard ℓ_2 norm for $\mathbf{y} \in \mathbb{R}^n$, and cond$(\mathbf{K})$ is the condition number of the matrix $\mathbf{K}$.
We can combine the above result with Lemma 5 to obtain the following bound on excess risk for minimum norm interpolating solutions to the kernel least squares problem:

Corollary 13 *The excess risk of the minimum norm interpolating solution to Problem* (3) *can be bounded as:*

$$\mathbb{E}_S \left[I[f_{S_i}^\dagger] - \inf_{f \in \mathcal{H}} I[f] \right] \leq C_1 \beta_1 + C_2 \beta_2$$

where β_1, β_2 are as defined previously.

Remark 14 (Underdetermined Linear Regression) *In the case of underdetermined linear regression, ie, linear regression where the dimensionality is larger than the number of samples in the training set, we can prove a version of Theorem 12 with* $\beta_1 = \mathbb{E}_S\left[\left\|\mathbf{X}^\dagger\right\|_{op}\|\mathbf{y}\|\right]$ *and* $\beta_2 = \mathbb{E}_S\left[\left\|\mathbf{X}^\dagger\right\|_{op}^2\|\mathbf{y}\|^2\right]$. *Due to space constraints we present the proof of the results in the linear regression case in Appendix B.*

The proofs of the kernel least squares results are given in the next sections. We first provide some comments. First, we can compare the above results with stability bound for penalized ERM, see Remark 3. Penalized ERM has a strong stability guarantee in terms of a uniform stability bound which turns out to be inversely proportional to the regularization parameter λ and the number of points n [3]. However, this estimate becomes vacuous as $\lambda \to 0$. For minimum norm solution we can only establish average stability. This is to be expected since one can expect worse case scenarios where the minimum norm is arbitrarily large for instance when $n \approx d$. This leads to a second observation, namely, that a different limit can be considered taking both the dimensionality of the data and the number of training points going to infinity. This is a classical setting in statistics which allows to use results from random matrix theory [18]. In particular, for linear kernels the behavior of the smallest eigenvalue of the kernel matrix can be characterized in this asymptotic limit. Here the dimension of the data coincides with the number of parameters in the model. Interestingly, analogous results can also be given for more general kernels [19] where the asymptotics are taken with respect to the number and dimensionality of the data (that is n and d). These results predict a double descent curve for the condition number as found in pracrice, see Figure 1.
Finally, we can compare this situation with observations [17] on the condition number of random kernel matrices and with results on the properties of minimum norm solutions. Recent papers consider linear models (kernels) and asymptotic regimes as discussed above, see e.g. [13] and references therein. The case of kernel based estimators is considered for example in [15, 16, 14]. Compared with these results our bound is simple and is derived from a stability argument, providing a natural link between numerical and statistical stability.

4.1 Key lemmas

In order to prove Theorem 12 we make use of the following lemmas to bound the CV_{loo} stability using the norms of the solutions. The first is standard, the second is our main estimate.

Lemma 15 *Under assumption* (9)*, for all* $i = 1 \dots, n$*, it holds that*

$$\mathbb{E}_S[V(f^\dagger_{S_i}, z_i) - V(f^\dagger_S, z_i)] \leq 2M\kappa\mathbb{E}_S\left[\left\|f^\dagger_S - f^\dagger_{S_i}\right\|_{\mathcal{H}}\right] + \kappa^2\mathbb{E}_S\left[\left(\left\|f^\dagger_S\right\|_{\mathcal{H}} + \left\|f^\dagger_{S_i}\right\|_{\mathcal{H}}\right)\left\|f^\dagger_S - f^\dagger_{S_i}\right\|_{\mathcal{H}}\right]$$

Proof We begin, recalling that the square loss is locally Lipschitz, that is for all $y, a, a' \in \mathbb{R}$, with

$$|(y-a)^2 - (y-a')^2| \leq (2|y| + |a| + |a'|))|a - a'|.$$

If we apply this result to f, f' in a RKHS $\mathcal{H}$,

$$|(y - f(\mathbf{x}))^2 - (y - f'(\mathbf{x}))^2| \leq \kappa(2M + \kappa\left(\|f\|_{\mathcal{H}} + \|f'\|_{\mathcal{H}}\right))\|f - f'\|_{\mathcal{H}}.$$

using the basic properties of a RKHS that for all $f \in \mathcal{H}$

$$|f(\mathbf{x})| \leq \|f\|_\infty \leq \kappa\|f\|_{\mathcal{H}} \tag{11}$$

In particular, we can plug $f^\dagger_{S_i}$ and $f^\dagger_S$ into the above inequality, and the almost positivity of ERM [22] will allow us to drop the absolute value on the left hand side. Finally the desired result follows by taking the expectation over S. ■

Now that we have bounded the CV_{loo} stability using the norms, we can find a bound on the norms of the solutions to the kernel least squares problem. This is our main estimate.

Lemma 16 *Let $f_S^\dagger$ be ad defined in* (5) *and $\hat{f}_S$ be any other interpolating solution, then $\left\|f_S^\dagger\right\|_{\mathcal{H}} \leq \left\|\hat{f}_S\right\|_{\mathcal{H}}$, and $\left\|f_S^\dagger - f_{S_i}^\dagger\right\|_{\mathcal{H}} \leq \left\|\hat{f}_S - \hat{f}_{S_i}\right\|_{\mathcal{H}}$.*
Also for some absolute constant C

$$\left\|f_S^\dagger - f_{S_i}^\dagger\right\|_{\mathcal{H}} \leq C \times \left\|\mathbf{K}^{\frac{1}{2}}\right\|_{op} \left\|\mathbf{K}^\dagger\right\|_{op} \times cond(\mathbf{K}) \times \|\mathbf{y}\| \tag{12}$$

Putting together Lemmas 15, 16 we obtain theorem 12. In the following section we provide the proof of Lemma 16.

4.2 Proof of Lemma 16

Recalling Section 2.1, we let $f_S^\dagger(\mathbf{x}) = \sum_{i=1}^n c_{S,i} K(\mathbf{x}_i, \mathbf{x})$ where $\mathbf{c}_S = \mathbf{K}^\dagger \mathbf{y}$, and $\mathbf{K} \in \mathbb{R}^{n\times n}$ is the kernel matrix K on S. i.e. $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, and $\mathbf{y} \in \mathbb{R}^n$ is the vector $\mathbf{y} = [y_1 \dots y_n]^\top$.
Similarly, the coefficient vector for the minimum norm ERM solution to the problem over the leave one out dataset S_i is $\mathbf{c}_{S_i} = (\mathbf{K}_{S_i})^\dagger \mathbf{y}_i$ Where $\mathbf{y}_i = [y_1, \dots, 0, \dots y_n]^\top$ and $\mathbf{K}_{S_i}$ is the kernel matrix $\mathbf{K}$ with the i^{th} row and column set to zero, which is the kernel matrix for the leave one out training set.
We define $\mathbf{a} = [-K(\mathbf{x}_1, \mathbf{x}_1), \dots, -K(\mathbf{x}_n, \mathbf{x}_n)]^\top \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^n$ as a one-hot column vector with all zeros apart from the i^{th} component which is 1. Let $\mathbf{a}_* = \mathbf{a} + K(\mathbf{x}_i, \mathbf{x}_i)\mathbf{b}$. Then, we have:

$$\begin{aligned} \mathbf{K}_* &= \mathbf{K} + \mathbf{b}\mathbf{a}_*^\top \\ \mathbf{K}_{S_i} &= \mathbf{K}_* + \mathbf{a}\mathbf{b}^\top \end{aligned} \tag{13}$$

That is, we can write $\mathbf{K}_{S_i}$ as a rank-2 update to $\mathbf{K}$. This can be verified by simple algebra, and using the fact that K is a symmetric kernel.
Now we are interested in bounding $\left\|f_S^\dagger - f_{S_i}^\dagger\right\|_{\mathcal{H}}$. For a function $h(\cdot) = \sum_{i=1}^m p_i K(\mathbf{x}_i, \cdot) \in \mathcal{H}$ we have $\|h\|_{\mathcal{H}} = \sqrt{\mathbf{p}^\top \mathbf{K}\mathbf{p}} = ||\mathbf{K}^{\frac{1}{2}}\mathbf{p}||$. So we have:

$$\begin{aligned} \left\|f_S^\dagger - f_{S_i}^\dagger\right\|_{\mathcal{H}} &= ||\mathbf{K}^{\frac{1}{2}}(\mathbf{c}_S - \mathbf{c}_{S_i})|| \\ &= ||\mathbf{K}^{\frac{1}{2}}(\mathbf{K}^\dagger \mathbf{y} - (\mathbf{K}_{S_i})^\dagger \mathbf{y}_i)|| \\ &= ||\mathbf{K}^{\frac{1}{2}}(\mathbf{K}^\dagger \mathbf{y} - (\mathbf{K}_{S_i})^\dagger \mathbf{y} + y_i(\mathbf{K}_{S_i})^\dagger \mathbf{b})|| \\ &\leq ||\mathbf{K}^{\frac{1}{2}}||_{op} \times ||\mathbf{K}^\dagger - (\mathbf{K}_{S_i})^\dagger||_{op} \times ||\mathbf{y}|| \end{aligned} \tag{14}$$

Here we make use of the fact that $(\mathbf{K}_{S_i})^\dagger \mathbf{b} = \mathbf{0}$.
If $\mathbf{K}$ has full rank (as in Remark 2), we see that $\mathbf{b}$ lies in the column space of $\mathbf{K}$ and $\mathbf{a}_*$ lies in the column space of $\mathbf{K}^\top$. Furthermore, $\beta_* = 1 + \mathbf{a}_*^\top \mathbf{K}^\dagger \mathbf{b} = 1 + \mathbf{a}^\top \mathbf{K}^\dagger \mathbf{b} + K(\mathbf{x}_i, \mathbf{x}_i)\mathbf{b}^\top \mathbf{K}^\dagger \mathbf{b} = \mathbf{K}_{ii}(\mathbf{K}^\dagger)_{ii} \neq 0$. Using equation 2.2 of [24] we obtain:

$$\begin{aligned} \mathbf{K}_*^\dagger &= \mathbf{K}^\dagger - (\mathbf{K}_{ii}(\mathbf{K}^\dagger)_{ii})^{-1}\mathbf{K}^\dagger \mathbf{b}\mathbf{a}_*^\top \mathbf{K}^\dagger \\ &= \mathbf{K}^\dagger - (\mathbf{K}_{ii}(\mathbf{K}^\dagger)_{ii})^{-1}\mathbf{K}^\dagger \mathbf{b}\mathbf{a}^\top \mathbf{K}^\dagger - ((\mathbf{K}^\dagger)_{ii})^{-1}\mathbf{K}^\dagger \mathbf{b}\mathbf{b}^\top \mathbf{K}^\dagger \\ &= \mathbf{K}^\dagger + (\mathbf{K}_{ii}(\mathbf{K}^\dagger)_{ii})^{-1}\mathbf{K}^\dagger \mathbf{b}\mathbf{b}^\top - ((\mathbf{K}^\dagger)_{ii})^{-1}\mathbf{K}^\dagger \mathbf{b}\mathbf{b}^\top \mathbf{K}^\dagger \end{aligned} \tag{15}$$

Next, we see that since $\mathbf{K}_*$ has the same rank as $\mathbf{K}$, $\mathbf{a}$ lies in the column space of $\mathbf{K}_*$, and $\mathbf{b}$ lies in the column space of $\mathbf{K}_*^\top$. Furthermore $\beta = 1 + \mathbf{b}^\top \mathbf{K}_* \mathbf{a} = 0$. This means we can use Theorem 6 in [25] (equivalent to formula 2.1 in [24]) to obtain the expression for $(\mathbf{K}_{S_i})^\dagger$

$$(\mathbf{K}_{S_i})^\dagger = \mathbf{K}_*^\dagger - \mathbf{k}\mathbf{k}^\dagger \mathbf{K}_*^\dagger - \mathbf{K}_*^\dagger \mathbf{h}^\dagger \mathbf{h} + (\mathbf{k}^\dagger \mathbf{K}_*^\dagger \mathbf{h}^\dagger)\mathbf{k}\mathbf{h} \tag{16}$$

where $\mathbf{k} = \mathbf{K}_*^\dagger \mathbf{a}$, $\mathbf{h} = \mathbf{b}^\top \mathbf{K}_*^\dagger$ and $\mathbf{v}^\dagger = \frac{\mathbf{v}^\top}{||\mathbf{v}||^2}$ for any non-zero vector $\mathbf{v}$.

$$\begin{aligned}
(\mathbf{K}_{S_i})^\dagger - \mathbf{K}_*^\dagger &= (\mathbf{k}^\dagger \mathbf{K}_*^\dagger \mathbf{h}^\dagger)\mathbf{k}\mathbf{h} - \mathbf{k}\mathbf{k}^\dagger \mathbf{K}_*^\dagger - \mathbf{K}_*^\dagger \mathbf{h}^\dagger \mathbf{h} \\
&= \mathbf{a}^\top (\mathbf{K}_*^\dagger)^\top \mathbf{K}_*^\dagger (\mathbf{K}_*^\dagger)^\top \mathbf{b} \times \frac{\mathbf{k}\mathbf{h}}{||\mathbf{k}||^2 ||\mathbf{h}||^2} - \mathbf{k}\mathbf{k}^\dagger \mathbf{K}_*^\dagger - \mathbf{K}_*^\dagger \mathbf{h}^\dagger \mathbf{h} \\
\implies ||(\mathbf{K}_{S_i})^\dagger - \mathbf{K}_*^\dagger||_{op} &\leq \frac{|\mathbf{a}^\top (\mathbf{K}_*^\dagger)^\top \mathbf{K}_*^\dagger (\mathbf{K}_*^\dagger)^\top \mathbf{b}|}{||\mathbf{K}_*^\dagger \mathbf{a}||||\mathbf{b}^\top \mathbf{K}_*^\dagger||} + 2||\mathbf{K}_*^\dagger||_{op} \\
&\leq \frac{||\mathbf{K}_*^\dagger||_{op}||\mathbf{K}_*^\dagger \mathbf{a}||||\mathbf{b}^\top \mathbf{K}_*^\dagger||}{||\mathbf{K}_*^\dagger \mathbf{a}||||\mathbf{b}^\top \mathbf{K}_*^\dagger||} + 2||\mathbf{K}_*^\dagger||_{op} \\
&= 3||\mathbf{K}_*^\dagger||_{op}
\end{aligned} \tag{17}$$

Above, we use the fact that the operator norm of a rank 1 matrix is given by $||\mathbf{u}\mathbf{v}^\top||_{op} = ||\mathbf{u}|| \times ||\mathbf{v}||$
Putting the two parts together we obtain the bound on $\left\|(\mathbf{K}_{S_i})^\dagger - \mathbf{K}^\dagger\right\|_{op}$:

$$\begin{aligned}
||\mathbf{K}^\dagger - (\mathbf{K}_{S_i})^\dagger||_{op} &= ||\mathbf{K}^\dagger - \mathbf{K}_*^\dagger + \mathbf{K}_*^\dagger - (\mathbf{K}_{S_i})^\dagger||_{op} \\
&\leq 3||\mathbf{K}_*^\dagger||_{op} + ||\mathbf{K}^\dagger - \mathbf{K}_*^\dagger||_{op} \\
&\leq 3||\mathbf{K}^\dagger||_{op} + 4(\mathbf{K}_{ii}(\mathbf{K}^\dagger)_{ii})^{-1}||\mathbf{K}^\dagger||_{op} + 4((\mathbf{K}^\dagger)_{ii})^{-1}||\mathbf{K}^\dagger||_{op}^2 \\
&\leq ||\mathbf{K}^\dagger||_{op}(3 + 8||\mathbf{K}^\dagger||_{op}||\mathbf{K}||_{op})
\end{aligned} \tag{18}$$

The last step follows from $(\mathbf{K}_{ii})^{-1} \leq ||\mathbf{K}^\dagger||_{op}$ and $((\mathbf{K}^\dagger)_{ii})^{-1} \leq ||\mathbf{K}||_{op}$. We can plug this bound into (14) to obtain the desired result.
We now turn to the first part of our lemma. If we choose an interpolating solution other than the minimum norm solution, then the stability parameter will be larger than what we have obtained here. Let us choose an interpolating solution $\hat{f}_S$ with coefficient vector $\hat{\mathbf{c}}_S = \mathbf{K}^\dagger \mathbf{y} + (\mathbf{I} - \mathbf{K}^\dagger \mathbf{K})\mathbf{v}$ for any $\mathbf{v} \in \mathbb{R}^n$. Now we have:

$$\begin{aligned}
||\hat{f}_S - \hat{f}_{S_i}||_{\mathcal{H}} &= ||\mathbf{K}^{\frac{1}{2}}(\hat{\mathbf{c}}_S - \hat{\mathbf{c}}_{S_i})|| \\
&= ||\mathbf{K}^{\frac{1}{2}}[\mathbf{K}^\dagger \mathbf{y} + (\mathbf{I} - \mathbf{K}^\dagger \mathbf{K})\mathbf{v} - (\mathbf{K}_{S_i})^\dagger \mathbf{y}_i + (\mathbf{I} - (\mathbf{K}_{S_i})^\dagger \mathbf{K}_{S_i})\mathbf{v}]|| \\
&= ||\mathbf{K}^{\frac{1}{2}}[(\mathbf{K}^\dagger - (\mathbf{K}_{S_i})^\dagger)\mathbf{y} + ((\mathbf{K}_{S_i})^\dagger \mathbf{K}_{S_i} - \mathbf{K}^\dagger \mathbf{K})\mathbf{v} + y_i(\mathbf{K}_{S_i})^\dagger \mathbf{b}]|| \\
&\leq \left\|\mathbf{K}^{\frac{1}{2}}\right\|_{op} \times \left[||\mathbf{K}^\dagger - (\mathbf{K}_{S_i})^\dagger||_{op} \times ||\mathbf{y}|| + ||(\mathbf{K}_{S_i})^\dagger \mathbf{K}_{S_i} - \mathbf{K}^\dagger \mathbf{K}||_{op} \times ||\mathbf{v}||\right] \\
&\leq \left\|\mathbf{K}^{\frac{1}{2}}\right\|_{op} \times \left[||\mathbf{K}^\dagger||_{op} \times (3 + 8||\mathbf{K}||_{op}||\mathbf{K}^\dagger||_{op}) \times ||\mathbf{y}|| + ||\mathbf{v}||\right]
\end{aligned} \tag{19}$$

Here we use equations from List 2 of [24] to obtain $(\mathbf{K}_{S_i})^\dagger \mathbf{K}_{S_i} = \mathbf{K}_*^\dagger \mathbf{K}_* - \mathbf{u}\mathbf{u}^\dagger = \mathbf{K}^\dagger \mathbf{K} - \mathbf{u}\mathbf{u}^\dagger$, where $\mathbf{u} = \mathbf{K}_*^\dagger \mathbf{a}$.
Thus the minimum norm interpolating solution is the most stable.

5 Conclusions

In summary, optimization of crossvalidation stability minimizes the expected error in both the classical and the modern regime of ERM. In the classical regime, CV_{loo} stability implies generalization and consistency for $n \to \infty$. In the modern regime, as described in this paper, stability can account for the double descent curve in kernel interpolants [11] under appropriate distributional assumptions. It has been claimed that stability can also explain why maximum margin solutions in deep networks, trained under exponential-type loss functions, minimize the expected error [23]. The main contribution of this paper is deriving excess risk bounds via a stability argument. In the process, we show that among the infinite number of interpolating

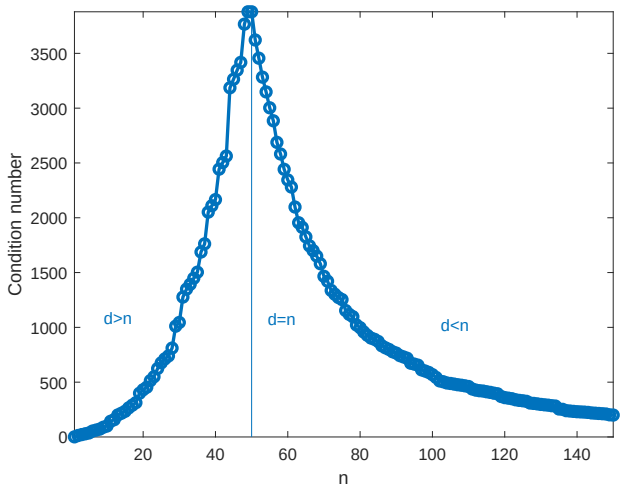

Figure 1: Typical double descent of the condition number (y axis) of a radial basis function kernel $K(x, x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$ built from a random data matrix distributed as $\mathcal{N}(0, 1)$: as in the linear case, the condition number is worse when $n = d$, better 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. From [17]

solutions, the one with minimal norm is the most stable both in a numerical and in statistical sense. This immediately yields information on solutions computed by gradient descent since they converge to minimum norm solutions in the case of "linear" kernel methods. Our approach is simple and combines basic stability results with matrix inequalities.

Acknowledgments 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.

References

[1] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.

[2] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. 2005.

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

[4] S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. Technical report TR-2002-03, University of Chicago, 2002.

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

[6] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, December 2010.

[7] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[8] Peter Bühlmann and Sara Van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*, volume 80. 2012.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[10] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[11] 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.

[12] 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.

[13] 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.

[14] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *CoRR*, abs/1906.11300, 2019.

[15] 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.

[16] 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.

[17] T. Poggio, G. Kur, and A. Banburski. Double descent in the condition number. Technical report, MIT Center for Brains Minds and Machines, 2019.

[18] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb. (N.S.)*, 72(114):4:457–483, 1967.

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

[20] Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1630–1638. Curran Associates, Inc., 2015.

[21] C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[22] 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.

[23] Tomaso Poggio. Stable foundations for learning. *Center for Brains, Minds and Machines (CBMM) Memo No. 103*, 2020.

[24] Jerzy K Baksalary, Oskar Maria Baksalary, and Götz Trenkler. A revisitation of formulae for the moore–penrose inverse of modified matrices. *Linear Algebra and Its Applications*, 372:207–224, 2003.

[25] Carl Meyer. Generalized inversion of modified matrices. *SIAM J. Applied Math*, 24:315–323, 1973.

A Excess Risk, Generalization, and Stability

We use the same notation as introduced in Section 2 for the various quantities considered in this section. That is in the supervised learning setup $V(f, z)$ is the loss incurred by hypothesis f on the sample z, and $I[f] = \mathbb{E}_z[V(f, z)]$ is the expected error of hypothesis f. Since we are interested in different forms of stability, we will consider learning problems over the original training set $S = \{z_1, z_2, \ldots, z_n\}$, the leave one out training set $S_i = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\}$, and the replace one training set $(S_i, z) = \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n, z\}$

A.1 Replace one and leave one out algorithmic stability

Similar to the definition of expected CV$_{loo}$ stability in equation (7) of the main paper, we say an algorithm is cross validation *replace one* stable (in expectation), denoted as CV$_{ro}$, if there exists $\beta_{ro} > 0$ such that

$$\mathbb{E}_{S,z}[V(f_S, z) - V(f_{(S_i,z)}, z)] \leq \beta_{ro}.$$

We can strengthen the above stability definition by introducing the notion of replace one algorithmic stability (in expectation) [3]. There exists $\alpha_{ro} >$ such that for all $i = 1, \ldots, n$,

$$\mathbb{E}_{S,z}[\left\|f_S - f_{(S_i,z)}\right\|_\infty] \leq \alpha_{ro}.$$

We make two observations:
First, if the loss is Lipschitz, that is if there exists $C_V > 0$ such that for all $f, f' \in \mathcal{H}$

$$\|V(f, z) - V(f', z)\| \leq C_V \|f - f'\|,$$

then replace one algorithmic stability implies CV$_{ro}$ stability with $\beta_{ro} = C_V \alpha_{ro}$. Moreover, the same result holds if the loss is locally Lipschitz and there exists $R > 0$, such that $\|f_S\|_\infty \leq R$ almost surely. In this latter case the Lipschitz constant will depend on R. Later, we illustrate this situation for the square loss.
Second, we have for all $i = 1, \ldots, n$, S and z,

$$\mathbb{E}_{S,z}[\left\|f_S - f_{(S_i,z)}\right\|_\infty] \leq \mathbb{E}_{S,z}[\|f_S - f_{S_i}\|_\infty] + \mathbb{E}_{S,z}[\left\|f_{(S_i,z)} - f_{S_i}\right\|_\infty].$$

This observation motivates the notion of leave one out algorithmic stability (in expectation) [3]]

$$\mathbb{E}_{S,z}[\|f_S - f_{S_i}\|_\infty] \leq \alpha_{loo}.$$

Clearly, leave one out algorithmic stability implies replace one algorithmic stability with $\alpha_{ro} = 2\alpha_{loo}$ and it implies also CV$_{ro}$ stability with $\beta_{ro} = 2C_V \alpha_{loo}$.

A.2 Excess Risk and CV$_{loo}$, CV$_{ro}$ Stability

We recall the statement of Lemma 5 in section 3 that bounds the excess risk using the CV$_{loo}$ stability of a solution.

Lemma 17 (Excess Risk & CV$_{loo}$ Stability) *For all $i = 1, \ldots, n$,*

$$\mathbb{E}_S[I[f_{S_i}] - \inf_{f \in \mathcal{H}} I[f]] \leq \mathbb{E}_S[V(f_{S_i}, z_i) - V(f_S, z_i)]. \tag{20}$$

In this section, two properties of ERM are useful, namely symmetry, and a form of unbiasedeness.

Symmetry. A key property of ERM is that it is *symmetric* with respect to the data set S, meaning that it does not depend on the order of the data in S.
A second property relates the expected ERM with the minimum of expected risk.

ERM Bias. The following inequality holds.

$$\mathbb{E}[[I_S[f_S]] - \min_{f\in\mathcal{H}} I[f] \leq 0. \tag{21}$$

To see this, note that

$$I_S[f_S] \leq I_S[f]$$

for all $f \in \mathcal{H}$ by definition of ERM, so that taking the expectation of both sides

$$\mathbb{E}_S[I_S[f_S]] \leq \mathbb{E}_S[I_S[f]] = I[f]$$

for all $f \in \mathcal{H}$. This implies

$$\mathbb{E}_S[I_S[f_S]] \leq \min_{f\in\mathcal{H}} I[f]$$

and hence (21) holds.

Remark 18 *Note that the same argument gives more generally that*

$$\mathbb{E}[\inf_{f\in\mathcal{H}}[I_S[f]] - \inf_{f\in\mathcal{H}} I[f] \leq 0. \tag{22}$$

Given the above premise, the proof of Lemma 5 is simple.

Proof [of Lemma 5] Adding and subtracting $\mathbb{E}_S[I_S[f_S]]$ from the expected excess risk we have that

$$\mathbb{E}_S[I[f_{S_i}] - \min_{f\in\mathcal{H}} I[f]] = \mathbb{E}_S[I[f_{S_i}] - I_S[f_S] + I_S[f_S] - \min_{f\in\mathcal{H}} I[f]], \tag{23}$$

and since $\mathbb{E}_S[I_S[f_S]] - \min_{f\in\mathcal{H}} I[f]]$ is less or equal than zero, see (22), then

$$\mathbb{E}_S[I[f_{S_i}] - \min_{f\in\mathcal{H}} I[f]] \leq \mathbb{E}_S[I[f_{S_i}] - I_S[f_S]]. \tag{24}$$

Moreover, for all $i = 1, \ldots, n$

$$\mathbb{E}_S[I[f_{S_i}]] = \mathbb{E}_S[\mathbb{E}_{z_i} V(f_{S_i}, z_i)] = \mathbb{E}_S[V(f_{S_i}, z_i)]$$

and

$$\mathbb{E}_S[I_S[f_S]] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_S[V(f_S, z_i)] = \mathbb{E}_S[V(f_S, z_i)].$$

Plugging these last two expressions in (24) and in (23) leads to (8). ■

We can prove a similar result relating excess risk with CV_{ro} stability.

Lemma 19 (Excess Risk & CV_{ro} Stability) *Given the above definitions, the following inequality holds for all* $i = 1, \ldots, n$,

$$\mathbb{E}_S[I[f_S] - \inf_{f\in\mathcal{H}} I[f]] \leq \mathbb{E}_S[I[f_S] - I_S[f_S]] = \mathbb{E}_{S,z}[V(f_S, z) - V(f_{(S_i,z)}, z)]. \tag{25}$$

Proof The first inequality is clear from adding and subtracting $I_S[f_S]$ from the expected risk $I[f_S]$ we have that

$$\mathbb{E}_S[I[f_S] - \min_{f\in\mathcal{H}} I[f]] = \mathbb{E}_S[I[f_S] - I_S[f_S] + I_S[f_S] - \min_{f\in\mathcal{H}} I[f]],$$

and recalling (22). The main step in the proof is showing that for all $i = 1, \ldots, n$,

$$\mathbb{E}[I_S[f_S]] = \mathbb{E}[V(f_{(S_i,z)}, z)] \tag{26}$$

to be compared with the trivial equality, $\mathbb{E}[I_S[f_S] = \mathbb{E}[V(f_S, z_i)]$. To prove Equation (26), we have for all $i = 1, \ldots, n$,

$$\mathbb{E}_S[I_S[f_S]] = \mathbb{E}_{S,z}[\frac{1}{n}\sum_{i=1}^{n} V(f_S, z_i)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{S,z}[V(f_{(S_i,z)}, z)] = \mathbb{E}_{S,z}[V(f_{(S_i,z)}, z)]$$

where we used the fact that by the symmetry of the algorithm $\mathbb{E}_{S,z}[V(f_{(S_i,z)}, z)]$ is the same for all $i = 1, \ldots, n$. The proof is concluded noting that $\mathbb{E}_S[I[f_S]] = \mathbb{E}_{S,z}[V(f_S, z)]$. ■

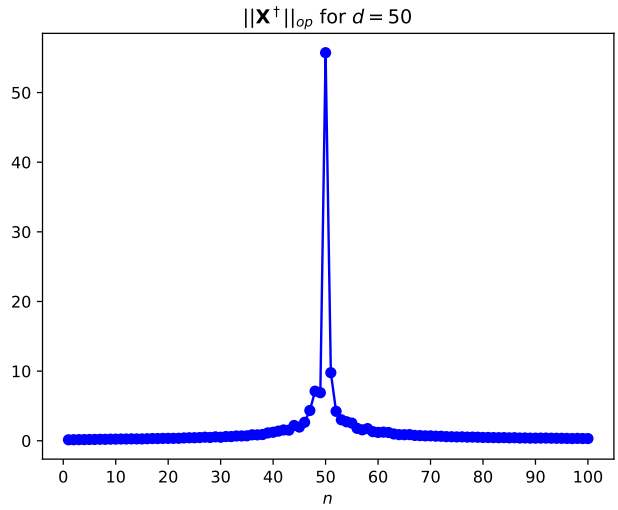

Figure 2: Typical double descent of the pseudoinverse norm (y axis) of a random data matrix distributed as $\mathcal{N}(0,1)$: the condition number is worse when $n = d$, better if $n > d$ (on the right of $n = d$) and also better if $n < d$ (on the left of $n = d$).. From [17]

B CV_{loo} Stabililty of Linear Regression

In this section we want to estimate the CV_{loo} stability of the minimum norm solution to the ERM problem in the linear regression case. This is the case outlined in Remark 14 of the main paper. In order to prove Remark 14, we only need to combine Lemma 15 with the linear regression analogue of Lemma 16. We state and prove that result in this section. This result predicts a double descent curve for the norm of the pseudoinverse as found in practice, see Figure 2.

Lemma 20 *Let $\mathbf{w}_S^\dagger$ be the minimum norm interpolating solution to the linear regression problem as defined in Remark 1 in the main paper and $\hat{\mathbf{w}}_S$ be any other interpolating solution, then $\left\|\mathbf{w}_S^\dagger\right\| \leq \|\hat{\mathbf{w}}_S\|$, and $\left\|\mathbf{w}_S^\dagger - \mathbf{w}_{S_i}^\dagger\right\| \leq \|\hat{\mathbf{w}}_S - \hat{\mathbf{w}}_{S_i}\|$. Also*

$$\left\|\mathbf{w}_S^\dagger - \mathbf{w}_{S_i}^\dagger\right\| \leq 3 \left\|\mathbf{X}^\dagger\right\|_{op} \times \|\mathbf{y}\| \tag{27}$$

As mentioned before in section 2.1 of the main paper, linear regression can be viewed as a case of the kernel regression problem where $\mathcal{H} = \mathbb{R}^d$, and the feature map Φ is the identity map. The inner product and norms considered in this case are also the usual Euclidean inner product and 2-norm for vectors in $\mathbb{R}^d$. The notation $\|\cdot\|$ denotes the Euclidean norm for vectors both in $\mathbb{R}^d$ and $\mathbb{R}^n$. The usage of the norm should be clear from the context. Also, $\|\mathbf{A}\|_{op}$ is the left operator norm for a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, that is $\|\mathbf{A}\|_{op} = \sup_{\mathbf{y}\in\mathbb{R}^n, ||\mathbf{y}||=1} ||\mathbf{y}^\top \mathbf{A}||$. We have n samples in the training set for a linear regression problem, $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. We collect all the samples into a single matrix/vector $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, and $\mathbf{y} = [y_1 y_2 \dots y_n]^\top \in \mathbb{R}^n$. Then any interpolating ERM solution $\mathbf{w}_S$ satisfies the linear equation

$$\mathbf{w}_S^\top \mathbf{X} = \mathbf{y}^\top \tag{28}$$

If we pick the minimum norm solution, then $\mathbf{w}_S^\dagger$ is given by

$$(\mathbf{w}_S^\dagger)^\top = \mathbf{y}^\top \mathbf{X}^\dagger. \tag{29}$$

If we consider the leave one out training set S_i we can find the minimum norm ERM solution for $\mathbf{X}_i = [\mathbf{x}_1 \dots \mathbf{0} \dots \mathbf{x}_n]$ and $\mathbf{y}_i = [y_1 \dots 0 \dots y_n]^\top$ as

$$(\mathbf{w}_{S_i}^\dagger)^\top = \mathbf{y}_i^\top (\mathbf{X}_i)^\dagger. \tag{30}$$

We can write $\mathbf{X}_i$ as:

$$\mathbf{X}_i = \mathbf{X} + \mathbf{a}\mathbf{b}^\top \tag{31}$$

where $\mathbf{a} \in \mathbb{R}^d$ is a column vector representing the additive change to the i^{th} column, i.e, $\mathbf{a} = -\mathbf{x}_i$, and $\mathbf{b} \in \mathbb{R}^{n\times 1}$ is the $i-$th element of the canonical basis in $\mathbb{R}^n$ (all the coefficients are zero but the $i-$th which is one). Thus $\mathbf{a}\mathbf{b}^\top$ is a $d \times n$ matrix composed of all zeros apart from the i^{th} column which is equal to $\mathbf{a}$.
We can also write $\mathbf{y}_i$ as:

$$\mathbf{y}_i = \mathbf{y} - y_i \mathbf{b} \tag{32}$$

Now per Lemma 15 we are interested in bounding the quantity $||\mathbf{w}_{S_i}^\dagger - \mathbf{w}_S^\dagger|| = ||(\mathbf{w}_{S_i}^\dagger)^\top - (\mathbf{w}_S^\dagger)^\top||$. This simplifies to:

$$\begin{aligned}
||\mathbf{w}_{S_i}^\dagger - \mathbf{w}_S^\dagger|| &= ||\mathbf{y}_i^\top (\mathbf{X}_i)^\dagger - \mathbf{y}^\top \mathbf{X}^\dagger|| \\
&= ||(\mathbf{y}^\top - y_i \mathbf{b}^\top)(\mathbf{X}_i)^\dagger - \mathbf{y}^\top \mathbf{X}^\dagger|| \\
&= ||\mathbf{y}^\top ((\mathbf{X}_i)^\dagger - \mathbf{X}^\dagger) + y_i \mathbf{b}^\top (\mathbf{X}_i)^\dagger|| \\
&= ||\mathbf{y}^\top ((\mathbf{X}_i)^\dagger - \mathbf{X}^\dagger)|| \\
&\leq \left\|(\mathbf{X}_i)^\dagger - \mathbf{X}^\dagger\right\|_{op} \|\mathbf{y}\|
\end{aligned} \tag{33}$$

We also make use of the fact that $\mathbf{b}^\top (\mathbf{X}_i)^\dagger = \mathbf{0}$. We can thus get a bound on the CV_{loo} stability of the minimum norm interpolating linear regression if we have a bound on $||(\mathbf{X}_i)^\dagger - \mathbf{X}^\dagger||_{op}$
We use an old formula [25, 24] to compute $(\mathbf{X}_i)^\dagger$ from $\mathbf{X}^\dagger$. We use the development of pseudo-inverses of perturbed matrices in [25]. We see that $\mathbf{a} = -\mathbf{x}_i$ is a vector in the column space of $\mathbf{X}$ and $\mathbf{b}$ is in the range space of $\mathbf{X}^T$ (provided $\mathbf{X}$ has full column rank), with $\beta = 1 + \mathbf{b}^\top \mathbf{X}^\dagger \mathbf{a} = 1 - \mathbf{b}^\top \mathbf{X}^\dagger \mathbf{x}_i = 0$. This means we can use Theorem 6 in [25] (equivalent to formula 2.1 in [24]) to obtain the expression for $(\mathbf{X}_i)^\dagger$

$$(\mathbf{X}_i)^\dagger = \mathbf{X}^\dagger - \mathbf{k}\mathbf{k}^\dagger \mathbf{X}^\dagger - \mathbf{X}^\dagger \mathbf{h}^\dagger \mathbf{h} + (\mathbf{k}^\dagger \mathbf{X}^\dagger \mathbf{h}^\dagger)\mathbf{k}\mathbf{h} \tag{34}$$

where

$$\begin{aligned}
\mathbf{k} &= \mathbf{X}^\dagger \mathbf{a} \\
\mathbf{h} &= \mathbf{b}^\top \mathbf{X}^\dagger
\end{aligned} \tag{35}$$

and $\mathbf{v}^\dagger = \frac{\mathbf{v}^\top}{||\mathbf{v}||^2}$ for any non-zero vector $\mathbf{v}$.

$$\begin{aligned}
(\mathbf{X}_i)^\dagger - \mathbf{X}^\dagger &= (\mathbf{k}^\dagger \mathbf{X}^\dagger \mathbf{h}^\dagger)\mathbf{k}\mathbf{h} - \mathbf{k}\mathbf{k}^\dagger \mathbf{X}^\dagger - \mathbf{X}^\dagger \mathbf{h}^\dagger \mathbf{h} \\
&= \mathbf{a}^\top (\mathbf{X}^\dagger)^\top \mathbf{X}^\dagger (\mathbf{X}^\dagger)^\top \mathbf{b} \times \frac{\mathbf{k}\mathbf{h}}{||\mathbf{k}||^2 ||\mathbf{h}||^2} - \mathbf{k}\mathbf{k}^\dagger \mathbf{X}^\dagger - \mathbf{X}^\dagger \mathbf{h}^\dagger \mathbf{h} \\
\implies ||(\mathbf{X}_i)^\dagger - \mathbf{X}^\dagger||_{op} &\leq \frac{|\mathbf{a}^\top (\mathbf{X}^\dagger)^\top \mathbf{X}^\dagger (\mathbf{X}^\dagger)^\top \mathbf{b}|}{||\mathbf{X}^\dagger \mathbf{a}||||\mathbf{b}^\top \mathbf{X}^\dagger||} + 2||\mathbf{X}^\dagger||_{op} \\
&\leq \frac{||\mathbf{X}^\dagger||_{op}||\mathbf{X}^\dagger \mathbf{a}||||\mathbf{b}^\top \mathbf{X}^\dagger||}{||\mathbf{X}^\dagger \mathbf{a}||||\mathbf{b}^\top \mathbf{X}^\dagger||} + 2||\mathbf{X}^\dagger||_{op} \\
&= 3||\mathbf{X}^\dagger||_{op}
\end{aligned} \tag{36}$$

The above set of inequalities follows from the fact that the operator norm of a rank 1 matrix is given by $||\mathbf{u}\mathbf{v}^\top||_{op} = ||\mathbf{u}|| \times ||\mathbf{v}||$

Finally we can get a bound on the stability as:

$$||\mathbf{w}_{S_i}^{\dagger} - \mathbf{w}_S^{\dagger}|| \leq 3||\mathbf{X}^{\dagger}||_{op}||\mathbf{y}|| \tag{37}$$

Now let us turn to the first part of our lemma. If our algorithm picks an interpolating solution other than $(\mathbf{w}_S^{\dagger})^{\top} = \mathbf{y}^{\top}\mathbf{X}^{\dagger}$, then the stability parameter will be larger than what we have obtained here. Let us say the interpolating solution is $\hat{\mathbf{w}}_S^{\top} = \mathbf{y}^{\top}\mathbf{X}^{\dagger} + \mathbf{v}^{\top}(\mathbf{I} - \mathbf{X}\mathbf{X}^{\dagger})$ for any $\mathbf{v} \in \mathbb{R}^d$. Now we have:

$$\begin{aligned}
||\hat{\mathbf{w}}_{S_i} - \hat{\mathbf{w}}_S|| &= ||\mathbf{y}_i^{\top}(\mathbf{X}_i)^{\dagger} + \mathbf{v}^{\top}(\mathbf{I} - \mathbf{X}_i(\mathbf{X}_i)^{\dagger}) - \mathbf{y}^{\top}\mathbf{X}^{\dagger} - \mathbf{v}^{\top}(\mathbf{I} - \mathbf{X}\mathbf{X}^{\dagger})|| \\
&\leq ||(\mathbf{X}_i)^{\dagger} - \mathbf{X}^{\dagger}||_{op}||\mathbf{y}|| + ||\mathbf{X}_i(\mathbf{X}_i)^{\dagger} - \mathbf{X}\mathbf{X}^{\dagger}||_{op}||\mathbf{v}|| \\
&\leq 3||\mathbf{X}^{\dagger}||_{op}||\mathbf{y}|| + ||\mathbf{v}|| \times 1
\end{aligned}$$

Here we use the fact from List 2 of [24] that $\mathbf{X}_i(\mathbf{X}_i)^{\dagger} = \mathbf{X}\mathbf{X}^{\dagger} - \mathbf{h}^{\dagger}\mathbf{h}$. Thus the minimum norm interpolating solution is the most stable among the ERM solutions.