For interpolating kernel machines, minimizing the norm of the ERM solution maximizes stability

Akshay Rangamani [*1,2], Lorenzo Rosasco [†1,3], and Tomaso Poggio [‡1,2]

[1]Center for Brains, Minds, and Machines, MIT
[2]McGovern Institute for Brain Research, MIT
[3]MaLGa, DIBRIS, University of Genova, Instituto Italiano di Tecnologia

Abstract

In this paper we study kernel ridge-less regression, including the case of interpolating solutions. We prove that maximizing the leave-one-out (CV_{loo}) stability minimizes the expected error. Further, we also prove that the minimum norm solution – to which gradient algorithms are known to converge – is the most stable solution. More precisely, we show that the minimum norm interpolating solution minimizes a bound on CV_{loo} stability, which in turn is controlled by the smallest singular value, hence the condition number, of the empirical kernel matrix. These quantities can be characterized in the asymptotic regime where both the dimension (d) and cardinality (n) of the data go to infinity (with $\frac{n}{d} \rightarrow \gamma$ as $d, n \rightarrow \infty$). Our results suggest that the property of CV_{loo} *stability* of the learning algorithm with respect to perturbations of the training set may provide a more general framework than the classical theory of Empirical Risk Minimization (ERM). While ERM was developed to deal with the *classical regime* in which the architecture of the learning network is fixed and n, the number of training examples, goes to infinity, the *modern regime* focuses on interpolating regressors and overparamerized models, when both d and n go to infinity. Since the stability framework is known to be equivalent to the classical theory in the classical regime, our results here suggest that it may be interesting to extend it beyond kernel regression to other overparametrized algorithms such as deep networks.

1 Introduction

Statistical learning theory studies the learning properties of machine learning algorithms, and more fundamentally, the conditions under which learning from finite data is possible. In this context, classical learning 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 (Boucheron et al., 2005; Shalev-Shwartz & Ben-David, 2014). Another, more recent, approach is based on defining suitable notions of stability with respect to perturbation of the data (Bousquet & Elisseeff, 2001; Kutin & Niyogi, 2002). 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 analyzed, depending on the data perturbation and considered error metric (Kutin & Niyogi, 2002). Interestingly, the stability and complexity approaches to characterizing the *learnability* of problems are not at odds with each other, and have been be shown to be equivalent in the classical framework, as shown in Poggio et al. (2004) and Shalev-Shwartz et al. (2010).
In modern machine learning overparameterized models, with a number of parameters larger than the size of the training data, have now become increasingly common. The ability of these models to generalize is well

[*] arangam@mit.edu
[†] lrosasco@mit.edu
[‡] tp@csail.mit.edu

explained by classical statistical learning theory as long as some form of regularization is used in the training process (Steinwart & Christmann, 2008; Bühlmann & Van De Geer, 2011). However, it was recently shown - first for deep networks (Zhang et al., 2017), and more recently for kernel methods (Belkin et al., 2019, 2018) - that learning is possible in the absence of regularization, i.e., when perfectly fitting/interpolating the data. Recent work in statistical learning theory has tried to find theoretical ground for this empirical finding. Since learning using models that interpolate is not exclusive to deep neural networks, we study generalization in the presence of interpolation in the case of kernel methods, with linear models as a special case.

Our Contributions:

- We characterize the generalization properties of possibly interpolating kernel ridge-less regression using a stability approach. While the (uniform) stability properties of regularized kernel methods are well known (Bousquet & Elisseeff, 2001), we study unregularized ("ridgeless") regression problems.
- We obtain an upper bound on the leave-one-out stability β_{CV} (defined later) of solutions to the kernel least squares problem, and show that this upper bound is minimized by the minimum norm interpolating solution. This also means that among all interpolating solutions, the minimum norm solution has the best test error[1]. In particular, the same conclusion is also true for gradient descent and stochastic gradient descent, since these algorithms converge to the minimum norm solution in the setting we consider, see e.g. Rosasco & Villa (2015).
- Our stability bounds show that the average stability of the minimum norm solution can be controlled by the minimum eigenvalue of the empirical kernel matrix. It is well known that the numerical stability of the least squares solution is governed by the condition number of the associated kernel matrix which is closely related to the minimum eigenvalue (see the discussion of why overparametrization is "good" in Poggio et al. (2019)). Our results show that the condition number also controls stability (and hence, test error) in a statistical sense.

Paper Outline: The rest of the paper is organized as follows. In section 2, we introduce basic ideas in statistical learning and empirical risk minimization, as well as the notation used in the rest of the paper. In section 3, we briefly recall some definitions of stability and their connection to test error. In this section we also provide a brief discussion about the promise of stability as a framework for the analysis of learning algorithms. In section 4, we present our main results on the stability of kernel least squares. The proof of our theorem is developed in section 6, where we also show that the minimum norm solutions minimize an upper bound on the stability. In section 5 we discuss our results in the context of recent work on high dimensional regression. We support our theoretical results with simulations in section 7 and conclude in section 8.

2 Statistical Learning and Empirical Risk Minimization

We begin by recalling the basic ideas in statistical learning theory. In this setting, X is the input space, Y is the output space, and there is an unknown probability distribution μ on $Z = X \times Y$. 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. $y_{new} \approx f_S(\mathbf{x}_{new})$.
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, that measures how well a function performs on a data point. We define a hypothesis space $\mathcal{H} \subseteq \mathcal{F}$ where algorithms search for solutions. With the above notation, the *expected risk* 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 μ. In this setting, statistical learning can be seen as the problem of finding an approximate minimizer of the expected risk

[1] This holds unless additional information is available, for instance about the data.

given a training set S. A classical approach to derive an approximate solution is empirical risk minimization (ERM) where we minimize the empirical risk $I_S[f] = \frac{1}{n}\sum_{i=1}^n V(f, z_i)$.
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 excess risk is easily bounded by the expected generalization error (see Lemma 5).

2.1 Kernel Least Squares and Minimum Norm Solution

The focus in this paper is kernel least squares. We assume the loss function V is the square loss, that is, $V(f,z) = (y - f(\mathbf{x}))^2$. The hypothesis space is assumed to be a reproducing kernel Hilbert space, defined by a positive definite kernel $K : X \times X \to \mathbb{R}$ with $\Phi : X \to \mathcal{H}$ an associated feature map, 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 $\mathbf{w} \in \mathcal{H}$ such that $f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x})\rangle_{\mathcal{H}}$ for all $\mathbf{x} \in X$.
The ERM problem typically has multiple solutions, one of which is the minimum norm solution:

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

where $\|\cdot\|_{\mathcal{H}}$ is the norm in $\mathcal{H}$. The minimum norm solution is unique and satisfies a representer theorem, for all $\mathbf{x} \in X$:

$$f_S^\dagger(\mathbf{x}) = \sum_{i=1}^n K(\mathbf{x}, \mathbf{x}_i)\mathbf{c}_{S,i}, \qquad \mathbf{c}_S = \mathbf{K}^\dagger \mathbf{y} \tag{2}$$

where $\mathbf{c}_S = (\mathbf{c}_{S,1}, \ldots, \mathbf{c}_{S,n}), \mathbf{y} = (y_1 \ldots y_n) \in \mathbb{R}^n$, $\mathbf{K}$ 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 $\mathbf{K}$. Since the input points are typically distinct, it is possible to show that for many kernels one can replace $\mathbf{K}^\dagger$ by $\mathbf{K}^{-1}$ (see Remark 2). Note that invertibility is necessary and sufficient for interpolation: if $\mathbf{K}$ is invertible, $f_S^\dagger(\mathbf{x}_i) = y_i$ for all $i = 1, \ldots, n$, in which case the training error in (1) is zero.
An alternative to using the explicit representation of the kernel matrix $\mathbf{K}$ is to represent linear functions in the RKHS as $f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x})\rangle$ for $\mathbf{w} \in \mathcal{H}$. If we collect the RKHS features of the data $\Phi(\mathbf{x}_i)$ into the rows of a linear operator $\mathbf{X} : \mathcal{H} \to \mathbb{R}^n$, then we can write the Kernel least square problem as $\min_{\mathbf{w}\in\mathcal{H}} ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2$. All interpolating solutions to this problem are of the form $\hat{\mathbf{w}}_S = \mathbf{X}^\dagger \mathbf{y} + (\mathbf{I}_{\mathcal{H}} - \mathbf{X}^\dagger\mathbf{X})\mathbf{v}$ for any $\mathbf{v} \in \mathcal{H}$. The relationship between the kernel matrix $\mathbf{K}$ and the operator $\mathbf{X}$ is $\mathbf{K} = \mathbf{X}\mathbf{X}^\top$.

Remark 1 (Pseudoinverse for underdetermined linear systems)
A simple, relevant example is the linear kernel where $f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}$, $\mathcal{H} = \mathbb{R}^d$ and Φ is the identity map. If the rank of $\mathbf{X} \in \mathbb{R}^{n\times d}$ is n, then any interpolating solution $\mathbf{w}_S$ satisfies $\mathbf{w}_S^\top \mathbf{x}_i = y_i$ for all $i = 1, \ldots, n$, and the minimum norm solution, also called Moore-Penrose solution, is given by $\mathbf{w}_S^\dagger = \mathbf{X}^\dagger\mathbf{y}$ where the pseudoinverse $\mathbf{X}^\dagger$ takes the form $\mathbf{X}^\dagger = (\mathbf{X}^\top\mathbf{X})^\dagger\mathbf{X}^\top$.

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 is 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 Micchelli (1986) 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 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 (1) 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 is given by $f_S^\lambda(\mathbf{x}) = \sum_{i=1}^n K(\mathbf{x}, \mathbf{x}_i)\mathbf{c}_i, \qquad \mathbf{c} = (\mathbf{K} + \lambda n \mathbf{I}_n)^{-1}\mathbf{y}$. In contrast to ERM 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 our introductory remarks by recalling the connection between minimum norm and the gradient descent.

Remark 4 (Minimum norm and gradient descent) *In our setting, it is well known that both batch and stochastic gradient (SGD) iterations converge to the minimum norm solution when multiple solutions exist, see e.g. Rosasco & Villa (2015). Thus, a study of the properties of the minimum norm solution explains the properties of the solution to which SGD converges. In particular, when ERM has multiple interpolating solutions, gradient descent converges to a solution minimizing a bound on stability, as we show next.*

3 Error Bounds via Stability

In this section, we present the definition of stability that we will be using in the paper, and discuss how stability may be a unifying framework for explaining learning in both the classical and modern regimes.
We first recall some basic results relating the learning and stability properties of Empirical Risk Minimization (ERM). Throughout the paper, we assume that ERM achieves a minimum, albeit the extension to almost minimizers is possible (Mukherjee et al., 2006) and important for exponential-type loss functions (Poggio, 2020). We do not require that a minimum exists for the expected risk. Since we will be considering leave-one-out stability in this section, we look at solutions to ERM 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 the definition laid out in Mukherjee et al. (2006) 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 $i = 1, \ldots, n$,

$$\mathbb{E}_S[V(f_{S_{-i}}, z_i) - V(f_S, z_i)] \leq \beta_{CV}. \tag{3}$$

Here $f_S, f_{S_{-i}}$ are the ERM solutions obtained on the full dataset and the leave one out dataset respectively. This definition is justified by the following result that bounds the excess risk of a learning algorithm by its average CV$_{loo}$ stability (Mukherjee et al., 2006; Shalev-Shwartz et al., 2010).

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

$$\mathbb{E}_S[I[f_{S_{-i}}] - \inf_{f \in \mathcal{H}} I[f]] \leq \mathbb{E}_S[V(f_{S_{-i}}, z_i) - V(f_S, z_i)]. \tag{4}$$

Remark 6 (Connection to other notions of stability) *Uniform stability, introduced by Bousquet & Elisseeff (2001), corresponds in 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 Kutin & Niyogi (2002) , Mukherjee et al. (2006).*

Lemma 5 is known and we recall the proof in Appendix A.2 for completeness. In Appendix A, we also discuss other definitions of stability and their connections to concepts in statistical learning theory like generalization and learnability.

3.1 The stability framework: a unifying principle for the classical and modern regimes

A milestone in classical learning theory was to formally show that appropriately restricting the hypothesis space – that is the space of functions represented by the learning machine – ensures consistency (and generalization) of ERM. The classical theory assumes that the hypothesis space is fixed while the number of training data n increases to infinity. Its basic results thus characterize the "classical" regime of $n > d$, where d is the number of parameters to be learned. The classical theory, however, cannot deal with what we call the "modern" regime, in which the network remains overparametrized $(n < d)$ when n grows. In this case the hypothesis space is not fixed: d increases as n increases. Different approaches that do not rely on the hypothesis space were developed already twenty years ago, motivated by learning algorithms that are not ERM , such as k-Nearest Neighbor. While trying to develop a theory that can deal with the classical *and* the modern regime, it seems natural to abandon the idea of the hypothesis space as the object of interest and focus instead on properties of supervised learning algorithms, which are maps from data sets to hypothesis functions. One can ask: *what property must the learning map L have for good generalization error?* The answer for a fixed hypothesis space is that CV$_{loo}$ stability is necessary and sufficient for generalization and consistency of ERM[2].

Building upon this observation, we conjecture that CV$_{loo}$ stability may be used to develop a unifying theory encompassing both the classical and the modern regime for ERM. In the classical regime generalization can take place provided $\beta_{CV} \to 0$ when $n \to \infty$; for ERM consistency follows from generalization. In the modern interpolatory regime, the generalization gap given by $\mathbb{E}_S[I[f_S] - I_S[f_S]]$ does not necessarily decrease to 0 as $n \to \infty$ since $I_S[f_S] = 0$, while we can have $I[f_S] > 0$ in general. However, for interpolating regressors, β_{CV} becomes a bound on the expected error. Thus the key claim of this unified approach is that *minimizing β_{CV} minimizes the expected generalization gap* and in particular minimizes the expected error in the modern regime. While this is satisfying conceptually, it is also important to spell out the implications of minimizing β_{CV} for ERM [3]. A natural answer is that minimizing β_{CV} in ERM may be equivalent to selecting the minimum norm solution. For the case of kernel regressors we show next that the minimum norm ERM interpolator indeed minimizes CV$_{loo}$ stability. It should be emphasized that this is an upper bound and we cannot expect the minimum norm solution to always yield the minimum expected error. In particular, better solution can be found when prior information is available (see Oravkin & Rebeschini (2021)). In addition, it remains an open question whether similar results hold for classifiers such as deep networks[4].

4 CV$_{loo}$ Stability of Kernel Least Squares

In this section we analyze the expected CV$_{loo}$ stability of solutions to the kernel least squares problem, and obtain a corresponding upper bound on their stability. We show that the upper bound on the expected CV$_{loo}$ stability is governed by the norm of the solutions in the case of interpolating solutions, and hence is the smallest for the minimum norm interpolating solution (1) .

As outlined in section 2.1, we consider a kernel least squares problem on a dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq (\mathbb{R}^d \times \mathbb{R})^n$. We use the linear parameterization in the RKHS $f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle_{\mathcal{H}}$, and collect the RKHS features of the data $\Phi(\mathbf{x}_i)$ into the rows of a linear operator $\mathbf{X} : \mathcal{H} \to \mathbb{R}^n$. We recall that all interpolating solutions are of the form $\hat{\mathbf{w}}_S = \mathbf{X}^\dagger \mathbf{y} + (\mathbf{I}_{\mathcal{H}} - \mathbf{X}^\dagger \mathbf{X})\mathbf{v}$ for some $\mathbf{v} \in \mathcal{H}$. Since we are interested in CV$_{loo}$ stability, we consider the same kernel least squares problem on the leave one out dataset S_{-i}. We replace the i^{th} row of $\mathbf{X}$ with $\mathbf{0}_{\mathcal{H}}$ to obtain the corresponding data operator $\mathbf{X}_{-i} : \mathcal{H} \to \mathbb{R}^n$ for the leave one out dataset. All solutions on the leave one out dataset S_{-i} can be written as $\hat{\mathbf{w}}_{S_{-i}} = (\mathbf{X}_{-i})^\dagger \mathbf{y}_{-i} + (\mathbf{I}_{\mathcal{H}} - \mathbf{X}_{-i}^\dagger \mathbf{X}_{-i})\mathbf{v}_{-i}$ for some $\mathbf{v}_{-i} \in \mathcal{H}$. We note that when $\mathbf{v} = \mathbf{0}_{\mathcal{H}}$ and $\mathbf{v}_{-i} = \mathbf{0}_{\mathcal{H}}$, we obtain the minimum norm interpolating solutions on the datasets S and S_{-i}.

Theorem 7 (Main Theorem) *Consider the kernel least squares problem where the inputs $\mathbf{x} \in \mathcal{H}$ and the outputs y are bounded, that is there exist $\kappa, M > 0$ such that*

$$||\mathbf{x}||_{\mathcal{H}}^2 \leq \kappa, \qquad |y| \leq M, \tag{5}$$

[2] LOO stability (see Poggio et al. (2004)) together with CV$_{loo}$ stability of the algorithm, both going to zero for $n \to \infty$ is sufficient for generalization for any supervised algorithm, including k-nearest neighbor and kernel machines.

[3] An argument may be made that while overparametrization makes sense for large but finite amounts of data, it should disappear for realistic learning machines as $n \to \infty$. If this does not happen the empirical loss will never converge to the expected loss, which seems a natural requirement, especially in a quasi-online setting.

[4] The results of course hold for deep RELU networks in the NTK regime, since they are then equivalent to kernel machines. Our results provide context for NTK analyses similar to that presented in Arora et al. (2019)

almost surely. Then for any interpolating solutions $\hat{f}_{S_{-i}}, \hat{f}_S$,

$$\mathbb{E}_S[V(\hat{f}_{S_{-i}}, z_i) - V(\hat{f}_S, z_i)] \leq C_1 \mathbb{E}_S\left[\beta_{CV}\right] + C_2 \mathbb{E}_S\left[\beta_{CV}^2\right] \tag{6}$$

Where $\beta_{CV} = ||\mathbf{X}^\dagger||_{op}||\mathbf{y}|| + 2||\mathbf{v} - \mathbf{v}_{-i}|| + ||\mathbf{v}_{-i}||$, and C_1, C_2 are absolute constants that do not depend on either d or n. This bound is minimized when $\mathbf{v} = \mathbf{v}_{-i} = \mathbf{0}_{\mathcal{H}}$, which corresponds to the minimum norm interpolating solutions $f_S^\dagger, f_{S_{-i}}^\dagger$. For the minimum norm solutions we have $\beta_{CV}^{min} = ||\mathbf{X}^\dagger||_{op}||\mathbf{y}||$.

In the above theorem $||\mathbf{X}^\dagger||_{op}$ refers to the operator norm of the pseudoinverse of the data operator $\mathbf{X}$, $||\mathbf{y}||$ refers to the (Euclidean) norm of $\mathbf{y} \in \mathbb{R}^n$.

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 8 *The excess risk of the minimum norm interpolating kernel least squares solution can be bounded as:*

$$\mathbb{E}_S\left[I[f_{S_{-i}}^\dagger] - \inf_{f \in \mathcal{H}} I[f]\right] \leq C_1 \mathbb{E}_S\left[||\mathbf{X}^\dagger||_{op}||\mathbf{y}||\right] + C_2 \mathbb{E}_S\left[||\mathbf{X}^\dagger||_{op}^2||\mathbf{y}||^2\right]$$

We provide the proof of Theorem 7 in section 6. In the next section we first offer some discussion of our results on stability, and put our results in the context of other recent results on interpolation in linear and kernel least squares problems.

5 Discussion and Related Work

In the previous section we obtained bounds on the CV$_{loo}$ stability of kernel least squares solutions and in particular of *interpolating* solutions. We established a bound on average stability for kernel least squares solutions, and show that this bound is minimized when the minimum norm ERM solution is selected. One of our key findings is the relationship between minimizing the norm of the ERM solution and minimizing a bound on stability. In this section we discuss our bound under different regimes of the sample size n and the dimensionality of the data d.

For the kernel least squares problem, interpolation occurs under mild conditions for different kernels. For instance, if the input data are all distinct, the inverse of the kernel matrix exists and for positive definite radial kernels interpolation is expected. For other kernels, such as the linear kernel, $d \geq n$ is needed for interpolation.

Asymptotic analysis: While our bounds hold for any finite d and n, it is worth understanding how they evolve under different regimes of n and d. For $n \to \infty$ (and d fixed), the smallest singular value of the kernel matrix $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]$ typically decreases with n. This means our bounds diverge and stability is lost. The classical approach here is to use regularized ERM (see Remark 3) corresponding to

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

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

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

Regularized ERM has a strong stability guarantee with a uniform stability bound (defined in appendix A.3) $\beta = O\left(\frac{1}{\lambda n}\right)$ which turns out to be inversely proportional to the regularization parameter λ and the sample size n (Bousquet & Elisseeff, 2001). Here the limit $n \to \infty$ implies asymptotic convergence to a zero stability gap.

In a setting that is common in statistics we can also consider how our bounds evolve as both the n and d go to infinity, but the ratio $\frac{n}{d} \to \gamma$ remains finite as $n, d \to \infty$. In this setting, it is possible to use results from random matrix theory (Marchenko & Pastur, 1967) to sketch the asymptotic limits of our bound for linear kernels under distributional assumptions on the data. Since $\left\|\mathbf{X}^\dagger\right\|_{op} = \frac{1}{\sigma_{\min}(\mathbf{X})}$, we can compute the asymptotic limit of our bound as $\left\|\mathbf{X}^\dagger\right\|_{op}^2 \|\mathbf{y}\|^2 = \frac{n}{d(1-\sqrt{\gamma})^2} = \frac{\gamma}{(1-\sqrt{\gamma})^2}$. Notice that the bound does not go to zero for $n \to \infty$ because in general the expected error cannot vanish (unless the classification labels are deterministic). Since the kernel matrix $\mathbf{K}$ is related to the data operator $\mathbf{X}$ as $\mathbf{K} = \mathbf{X}^\top \mathbf{X}$, we have that $\sigma_{\min}(\mathbf{K}) = \sigma_{\min}(\mathbf{X})^2$, and our bound can be written in terms of the minimum singular value of the kernel matrix rather than data operator.

Interestingly, properties analogous to the Marchenko-Pastur limit hold for more general kernels. Consider random matrices whose entries are $K(\mathbf{x}_i^\top \mathbf{x}_j)$ with i.i.d. vectors $\mathbf{x}_i$ in $\mathbb{R}^d$ with mean zero and unit variance. Assuming that the distribution of $\mathbf{x}_i$'s is sufficiently nice and f is sufficiently smooth, El Karoui (2010) showed that in the Marchenko–Pastur limit, the spectral distributions of kernel dot-product matrices $\mathbf{K}_{ij} = f(\frac{1}{d}\mathbf{x}_i^\top \mathbf{x}_j)$ behave as if f is linear. In fact, El Karoui showed that under mild conditions, the kernel matrix is asymptotically equivalent to a linear combination of the linear kernel matrix, the all-1's matrix, and the identity, and hence the limiting spectrum is Marcenko–Pastur [5].

However, we note that our results *do not* predict a double descent curve for kernels that are not linear dot product kernels (Poggio et al., 2019). We discuss this observation in more detail in Section 7.

Related Work: Recently, there has been a surge of interest in studying linear and kernel least squares models, since classical results focus on situations where constraints or penalties that prevent interpolation are added to the empirical risk. For example, high dimensional linear regression is considered in Mei & Montanari (2019); Hastie et al. (2019) from the perspective of asymptotic statistics. A non asymptotic approach is considered instead in Bartlett et al. (2019); Liang et al. (2019); Rakhlin & Zhai (2018) and Liang et al. (2020). In particular, the results in Bartlett et al. (2019) are the first to obtain convergence when the number of dimensions/parameters is fixed.

While these papers study upper and lower bounds on the risk of interpolating solutions to the linear and kernel least squares problem, ours are the first to be derived using stability arguments. While it might be possible to obtain tighter excess risk bounds through careful analysis of the minimum norm interpolant, our simple approach helps us establish a link between stability in a statistical and numerical sense. Of course, our result is in terms of an upper bound and since lower bounds do not yet exist and seem difficult to obtain, it is reasonable to be skeptical of its quantitative values. More relevant in our opinion is the qualitative statement that minimizing the norm of an interpolating solution has the effect of making its stability gap smaller and thus of minimizing its expected error. We also see this reflected in numerical simulations in section 7. Concurrent to our work, Liang & Recht (2021) study the classification problem using kernels and obtain a mistake bound for the minimum norm interpolating classifier. However they do not make the connection to CV$_{loo}$ stability.

Finally, we can compare our results with observations made in Poggio et al. (2019) on the condition number of random kernel matrices. The condition number of the empirical kernel matrix is known to control the numerical stability of the solution to a kernel least squares problem. Our results show that the statistical stability is also controlled by the minimum singular value of the kernel matrix (which is closely related to the condition number), providing a natural link between numerical and statistical stability.

[5] Remark 5.1 of Liang et al. (2020) observes that since the data is usually centered ($\sum_{i=1}^n \mathbf{x}_i = \mathbf{0}$), the spectrum of the the kernel matrix is close to the spectrum of the linear kernel.

6 Proof of Theorem 7

6.1 Key lemmas

In order to prove Theorem 7 we make use of the following lemmas to bound the CV$_{loo}$ stability using the norms and the difference of the solutions.

Lemma 9 *Under assumption* (5), *for all* $i = 1\ldots, n$, *it holds that*

$$\mathbb{E}_S[V(\hat{f}_{S_{-i}}, z_i) - V(\hat{f}_S, z_i)] \leq \mathbb{E}_S\left[\left(2M + \kappa\left(\left\|\hat{f}_S\right\|_{\mathcal{H}} + \left\|\hat{f}_{S_{-i}}\right\|_{\mathcal{H}}\right)\right) \times \kappa \left\|\hat{f}_S - \hat{f}_{S_{-i}}\right\|_{\mathcal{H}}\right]$$

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 $f \in \mathcal{H}$

$$|f(\mathbf{x})| \leq \|f\|_\infty \leq \kappa\,\|f\|_{\mathcal{H}} \tag{9}$$

In particular, we can plug $\hat{f}_{S_{-i}}$ and $\hat{f}_S$ into the above inequality, and the almost positivity of ERM (Mukherjee et al., 2006) 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 and the difference of the solutions, we can find a bound on the difference between the solutions to the kernel least squares problem. This is our main stability estimate.

Lemma 10 *Let* $\hat{f}_S, \hat{f}_{S_{-i}}$ *be any interpolating kernel least squares solutions on the full and leave one out datasets* (*as defined at the top of this section*), *then* $\left\|\hat{f}_S - \hat{f}_{S_{-i}}\right\|_{\mathcal{H}} \leq ||\mathbf{X}^\dagger||_{op}||\mathbf{y}|| + 2||\mathbf{v} - \mathbf{v}_{-i}|| + ||\mathbf{v}_{-i}||$. *This bound is minimized when* $\mathbf{v} = \mathbf{v}_{-i} = \mathbf{0}_{\mathcal{H}}$, *which corresponds to the minimum norm interpolating solutions* $f_S^\dagger, f_{S_{-i}}^\dagger$.
Also,

$$\left\|f_S^\dagger - f_{S_{-i}}^\dagger\right\| \leq \left\|\mathbf{X}^\dagger\right\|_{op}\|\mathbf{y}\| \tag{10}$$

Remark 11 (Zero training loss) *In Lemma 9 we use the locally Lipschitz property of the squared loss function to bound the leave one out stability in terms of the difference between the norms of the solutions. Under interpolating conditions, if we set the term* $V(\hat{f}_S, z_i) = 0$, *the leave one out stability reduces to* $\mathbb{E}_S\left[V(\hat{f}_{S_{-i}}, z_i) - V(\hat{f}_S, z_i)\right] = \mathbb{E}_S\left[V(\hat{f}_{S_{-i}}, z_i)\right] = \mathbb{E}_S[(\hat{f}_{S_{-i}}(\mathbf{x}_i) - y_i)^2] = \mathbb{E}_S[(\hat{f}_{S_{-i}}(\mathbf{x}_i) - \hat{f}_S(\mathbf{x}_i))^2] = \mathbb{E}_S[\langle \hat{f}_{S_{-i}}(\cdot) - \hat{f}_S(\cdot), K_{\mathbf{x}_i}(\cdot)\rangle^2] \leq \mathbb{E}_S\left[||\hat{f}_S - \hat{f}_{S_{-i}}||^2_{\mathcal{H}} \times \kappa^2\right]$. *We can plug in the bound from Lemma 10 to obtain similar qualitative and quantitative* (*up to constant factors*) *results as in Theorem 7.*

Since the minimum norm interpolating solutions minimize both $\left\|\hat{f}_S\right\|_{\mathcal{H}} + \left\|\hat{f}_{S_{-i}}\right\|_{\mathcal{H}}$ and $\left\|\hat{f}_S - \hat{f}_{S_{-i}}\right\|_{\mathcal{H}}$ (from lemmas 9, 10), we can put them together to prove theorem 7. In the following section we provide the proof of Lemma 10.

6.2 Proof of Lemma 10

We have n samples in the training set for a kernel least squares problem, $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. We consider the linear operator $\mathbf{X} = [\Phi(\mathbf{x}_1)^\top; \Phi(\mathbf{x}_2)^\top; \dots \Phi(\mathbf{x}_n)^\top]$ from $\mathcal{H}$ to $\mathbb{R}^n$ and vector of labels $\mathbf{y} = [y_1 y_2 \dots y_n]^\top \in \mathbb{R}^n$. For any $\mathbf{w} \in \mathcal{H}$, the operator evaluaties to $\mathbf{X}\mathbf{w} \in \mathbb{R}^n$, the i^{th} entry of which is given by $\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle_{\mathcal{H}}$. Then any ERM solution $\mathbf{w}_S$ satisfies the linear equation

$$\mathbf{X}\hat{\mathbf{w}}_S = \mathbf{y} \tag{11}$$

Any solution can be written as:

$$\hat{\mathbf{w}}_S = \mathbf{X}^\dagger \mathbf{y} + (\mathbf{I}_{\mathcal{H}} - \mathbf{X}^\dagger \mathbf{X})\mathbf{v} \tag{12}$$

If we consider the leave one out training set S_{-i} we can find the minimum norm ERM solution for $\mathbf{X}_{-i} = [\Phi(\mathbf{x}_1)^\top; \dots \mathbf{0}_{\mathcal{H}}^\top; \dots; \Phi(\mathbf{x}_n)^\top]$ and $\mathbf{y}_{-i} = [y_1 \dots 0 \dots y_n]^\top$ as

$$\hat{\mathbf{w}}_{S_{-i}} = (\mathbf{X}_{-i})^\dagger \mathbf{y}_{-i} + (\mathbf{I}_{\mathcal{H}} - \mathbf{X}_{-i}^\dagger \mathbf{X}_{-i})\mathbf{v}_{-i} \tag{13}$$

We can write $\mathbf{X}_{-i}$ as:

$$\mathbf{X}_{-i} = \mathbf{X} + \mathbf{a}\mathbf{b}^\top \tag{14}$$

where $\mathbf{b} \in \mathcal{H}$ is a vector representing the additive change to the i^{th} row, i.e, $\mathbf{b} = -\Phi(\mathbf{x}_i)$, and $\mathbf{a} = \mathbf{e}_i \in \mathbb{R}^n$ 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 linear operator from $\mathcal{H}$ to $\mathbb{R}^n$ that maps vectors in $\mathcal{H}$ to scaled versions of $\mathbf{a}$.

We also have $\mathbf{y}_{-i} = \mathbf{y} - y_i\mathbf{a}$. Now per Lemma 9 we are interested in bounding the quantity $||\hat{f}_{S_{-i}} - \hat{f}_S||_{\mathcal{H}} = ||\hat{\mathbf{w}}_{S_{-i}} - \hat{\mathbf{w}}_S||_{\mathcal{H}}$. This simplifies to:

$$\begin{aligned}
||\hat{\mathbf{w}}_{S_{-i}} - \hat{\mathbf{w}}_S||_{\mathcal{H}} &= ||\mathbf{X}_{-i}^\dagger \mathbf{y}_{-i} - \mathbf{X}^\dagger \mathbf{y} + \mathbf{v}_{-i} - \mathbf{v} + \mathbf{X}^\dagger \mathbf{X}\mathbf{v} - \mathbf{X}_{-i}^\dagger \mathbf{X}_{-i}\mathbf{v}_{-i}||_{\mathcal{H}} \\
&= ||(\mathbf{X}_{-i})^\dagger(\mathbf{y} - y_{-i}\mathbf{a}) - \mathbf{X}^\dagger \mathbf{y} + \mathbf{v}_{-i} - \mathbf{v} + \mathbf{X}^\dagger \mathbf{X}\mathbf{v} - \mathbf{X}_{-i}^\dagger \mathbf{X}_{-i}\mathbf{v}_{-i}||_{\mathcal{H}} \\
&= ||(\mathbf{X}_{-i}^\dagger - \mathbf{X}^\dagger)\mathbf{y} + y_{-i}\mathbf{X}_{-i}^\dagger \mathbf{a} + \mathbf{v}_{-i} - \mathbf{v} + \mathbf{X}^\dagger \mathbf{X}\mathbf{v} - \mathbf{X}_{-i}^\dagger \mathbf{X}_{-i}\mathbf{v}_{-i}|| \\
&= ||(\mathbf{X}_{-i}^\dagger - \mathbf{X}^\dagger)\mathbf{y} + \mathbf{v}_{-i} - \mathbf{v} + \mathbf{X}^\dagger \mathbf{X}\mathbf{v} - \mathbf{X}_{-i}^\dagger \mathbf{X}_{-i}\mathbf{v}_{-i}|| \\
&= ||(\mathbf{X}_{-i}^\dagger - \mathbf{X}^\dagger)\mathbf{y} + (\mathbf{I}_{\mathcal{H}} - \mathbf{X}^\dagger \mathbf{X})(\mathbf{v}_{-i} - \mathbf{v}) + (\mathbf{X}^\dagger \mathbf{X} - \mathbf{X}_{-i}^\dagger \mathbf{X}_{-i})\mathbf{v}_{-i}||
\end{aligned} \tag{15}$$

In the above equation we make use of the fact that $\mathbf{X}_{-i}^\dagger \mathbf{a} = \mathbf{0}_{\mathcal{H}}$. We use an old formula (Meyer, 1973; Baksalary et al., 2003) to compute $(\mathbf{X}_{-i})^\dagger$ from $\mathbf{X}^\dagger$. We use the development of pseudo-inverses of perturbed matrices in Meyer (1973). Since none of the theorems depend on the finite dimensionality of $\mathcal{H}$, we can use those results for linear operators. We see that $\mathbf{b} = -\Phi(\mathbf{x}_i)$ is a vector in the range of $\mathbf{X}^\top$ and $\mathbf{a}$ is in the range of $\mathbf{X}$ (provided $\mathbf{X}$ has rank n), with $\beta = 1 + \mathbf{b}^\top \mathbf{X}^\dagger \mathbf{a} = 1 - \Phi(\mathbf{x}_i)^\top \mathbf{X}^\dagger \mathbf{a} = 0$. This means we can use Theorem 6 in Meyer (1973) (equivalent to formula 2.1 in Baksalary et al. (2003)) to obtain the expression for $\mathbf{X}_{-i}^\dagger$

$$\mathbf{X}_{-i}^\dagger = \mathbf{X}^\dagger - \mathbf{k}\mathbf{k}^\dagger \mathbf{X}^\dagger - \mathbf{X}^\dagger \mathbf{h}^\dagger \mathbf{h} + (\mathbf{k}^\dagger \mathbf{X}^\dagger \mathbf{h}^\dagger)\mathbf{k}\mathbf{h} \tag{16}$$

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

$$\begin{aligned}
\mathbf{X}_{-i}^\dagger - \mathbf{X}^\dagger &= (\mathbf{k}^\dagger \mathbf{X}^\dagger \mathbf{h}^\dagger)\mathbf{k}\mathbf{h} - \mathbf{k}\mathbf{k}^\dagger \mathbf{X}^\dagger - \mathbf{X}^\dagger \mathbf{h}^\dagger \mathbf{h} \\
&= (\mathbf{k}^\dagger \mathbf{X}^\dagger \mathbf{a})\mathbf{k}\mathbf{a}^\top - \mathbf{k}\mathbf{k}^\dagger \mathbf{X}^\dagger - \mathbf{X}^\dagger \mathbf{a}\mathbf{a}^\top \\
&= (\mathbf{k}^\dagger \mathbf{k})\mathbf{k}\mathbf{a}^\top - \mathbf{k}\mathbf{a}^\top - \mathbf{k}\mathbf{k}^\dagger \mathbf{X}^\dagger \\
\implies ||\mathbf{X}_{-i}^\dagger - \mathbf{X}^\dagger||_{op} &= ||\mathbf{k}\mathbf{k}^\dagger \mathbf{X}^\dagger||_{op} \\
&\leq ||\mathbf{X}^\dagger||_{op}
\end{aligned} \tag{17}$$

The above set of inequalities follows from the fact that the operator norm of a rank 1 matrix is given by $||\mathbf{u}\mathbf{v}^\top||_{op} = ||\mathbf{u}|| \times ||\mathbf{v}||$, and by noticing that $\mathbf{k} = -\mathbf{b}$.
Also, from List 2 of Baksalary et al. (2003) we have that $\mathbf{X}_{-i}^\dagger \mathbf{X}_{-i} = \mathbf{X}^\dagger \mathbf{X} - \mathbf{k}\mathbf{k}^\dagger$.
Plugging in these calculations into equation 15 we get:

$$\begin{aligned} ||\hat{\mathbf{w}}_{S_{-i}} - \hat{\mathbf{w}}_S|| &= ||(\mathbf{X}_{-i}^\dagger - \mathbf{X}^\dagger)\mathbf{y} + (\mathbf{I}_\mathcal{H} - \mathbf{X}^\dagger\mathbf{X})(\mathbf{v}_{-i} - \mathbf{v}) - (\mathbf{X}^\dagger\mathbf{X} - \mathbf{X}_{-i}^\dagger\mathbf{X}_{-i})\mathbf{v}_{-i}|| \\ &\le ||\mathbf{X}^\dagger||_{op}||\mathbf{y}|| + ||\mathbf{I}_\mathcal{H} - \mathbf{X}^\dagger\mathbf{X}||_{op}||\mathbf{v} - \mathbf{v}_{-i}|| + ||\mathbf{k}\mathbf{k}^\dagger||_{op}||\mathbf{v}_{-i}|| \\ &\le ||\mathbf{X}^\dagger||_{op}||\mathbf{y}|| + 2||\mathbf{v} - \mathbf{v}_{-i}|| + ||\mathbf{v}_{-i}|| \end{aligned} \tag{18}$$

We see that the right hand side is minimized when $\mathbf{v} = \mathbf{v}_{-i} = \mathbf{0}_\mathcal{H}$. This concludes the proof of Lemma 10.

Remark 12 (Stability of the minimum norm solution) *We can perform a more careful analysis of the stability of the minimum norm solution by putting together equations* (15) *and* (17)*, with* $\mathbf{v} = \mathbf{v}_{-i} = \mathbf{0}_\mathcal{H}$ *to obtain the following bound:*

$$\left\|\mathbf{w}_{S_{-i}}^\dagger - \mathbf{w}_S^\dagger\right\| = \left\|-\mathbf{k}\mathbf{k}^\dagger\mathbf{X}^\dagger\mathbf{y}\right\| \le \left\|\mathbf{w}_S^\dagger\right\| \tag{19}$$

Putting this together with Lemma 9 and Lemma 5 we can see that the CV_{loo} *stability – and hence excess risk of the minimum norm solution to the kernel least squares problem – is bounded by the norm of the solution.*

7 Simulations

In this section we perform experiments to provide empirical evidence for our theoretical results. We first verify that the minimum norm interpolating solution maximizes stability among all interpolating solutions. Subsequently we show that the test error inversely correlates with the minimum singular value of the kernel matrix, and we finally verify that the norm of the minimum norm interpolating solution governs its test error. We will now see those experiments one by one.

Stability and norm of the solution: In order to illustrate that the minimum norm interpolating solution is the best performing interpolating solution we run a simple experiment on a linear regression problem. We synthetically generate data from a linear model $\mathbf{y} = \mathbf{X}\mathbf{w}$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is i.i.d $\mathcal{N}(0, 1)$. The dimension of the data is $d = 1000$ and there are $n = 200$ samples in the training dataset. We use a held out test dataset of 100 samples to measure the generalization performance. We compute interpolating solutions as described at the beginning of this section, $\hat{\mathbf{w}} = \mathbf{X}^\dagger\mathbf{y} + (\mathbf{I} - \mathbf{X}^\dagger\mathbf{X})\mathbf{v}$, using $\mathbf{v}$'s of different norms to compare the test error and CV$_{loo}$ stability. The results (averaged over 100 trials) are shown in Figure 1. In Figure 1a we can see that the training loss is 0 for all interpolating solutions, but the test MSE increases as $||\mathbf{v}||$ increases, with the minimum norm solution $\mathbf{w}^\dagger = \mathbf{X}^\dagger\mathbf{y}$ having the best performance. We also observe in Figure 1b that the CV$_{loo}$ stability, computed using the expression in (3), also follows a similar trend. From both plots in Figure 1 we can see that the minimum norm interpolating solution has the best stability as well as the best test error, as suggested by Theorem 7.

Test Error and $\frac{1}{\sigma_{\min}(\mathbf{K})}$: Our results also indicate that the bound on the CV$_{loo}$ stability (and hence the test error) of the minimum norm interpolating solution depends on the norm of the (pseudo) inverse of the empirical kernel matrix. In order to verify this using simulations, we consider a regression problem in which we learned the function $f(\mathbf{x}) = \exp(-2\,\|\mathbf{x}\|)$ using kernel least squares. We generated samples $\mathbf{x}_i \in \mathbb{R}^{20}$ and learned f using interpolating kernel "ridgeless" regression with a polynomial kernel of degree 2 as well as a radial basis function (RBF) kernel. The training dataset was generated by sampling $\mathbf{X} \in \mathbb{R}^{n \times d}$ $(d = 20, n = 200)$ i.i.d. from $\mathcal{N}(0, 1)$, and a held out test dataset of 100 samples was also generated in a similar fashion. The results of this simulation can be seen in Figure 2. In order to obtain RBF and polynomial kernel matrices with different singular values, we varied the size of the training dataset from 10 to 200 (Figure 2a, 2b). In both cases we observe that the log test MSE of the minimum norm interpolating solution is correlated with the log of the norm of the pseudoinverse of the empirical kernel matrix. This confirms our observation that the numerical stability and statistical stability of a kernel least squares problem are related through the smallest singular value of the kernel matrix.

We note that our results *do not* predict a double descent curve in the smallest singular value for kernels that are not linear dot product kernels (Poggio et al., 2019). In the case of linear dot product kernels, since the spectra of $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}\mathbf{X}^\top$ are the same, one can expect a double descent curve for the smallest singular value of the kernel matrix. This is not true for more general kernels. The expected loss should therefore also not show a double descent, except in the case of the linear kernel. This is also what we find empirically (see Figures 1a and 2a).

Test Error and norm of the solution: In order to show that the norm of the minimum norm solution to the kernel least squares problem also governs the stability and hence test error (remark 12), we performed an experiment on binary classification in which the fraction of random labels assigned was varied in order to increase the noise level of the problem. We generated data from two gaussian distributions with different means, ie from $\mathcal{N}(\pm 3 \times \mathbf{1}_{20}, \mathbf{I}_{20})$, and trained an RBF kernel on 200 training samples and observed the test error on a held out set of 100 samples. As we can see in Figure 3, the norm of the interpolating solution (red) and the test mean squared error (blue) both increase as the label noise increases. An intuitive explanation of the reason the norm grows is that the pseudoinverse of the data operator $\mathbf{X}^\dagger$ is effectively a high-pass filter that amplifies high-frequencies (more noise) in $\mathbf{y}$, and increases the norm of the minimum norm solution. We also expect the test error to grow as the label noise in the problem increases, since the minimum achievable error is atleast the label noise. Our results on the stability and hence test error of the minimum norm solution to the kernel least squares problem also capture this phenomenon.

8 Conclusions

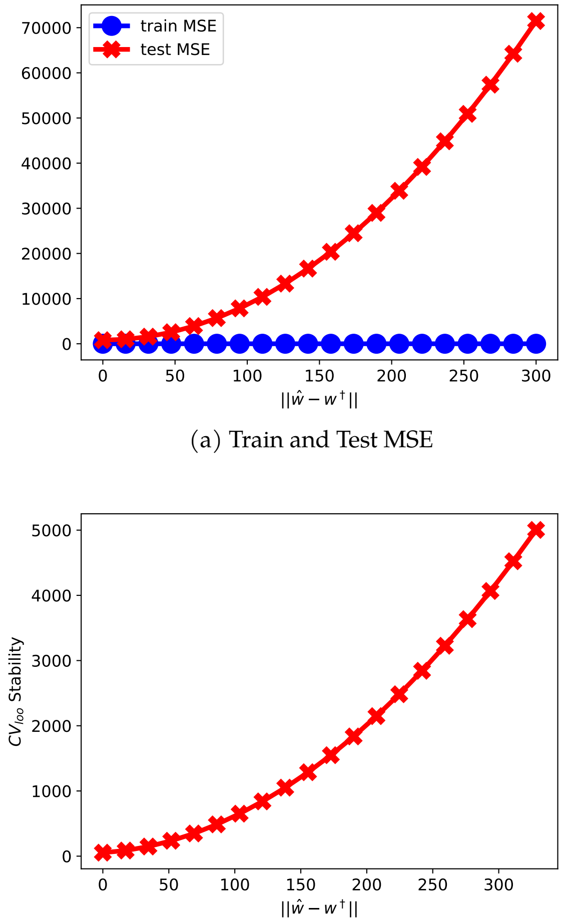
In summary, minimizing a bound on cross validation stability minimizes the expected error in both the classical and the modern regime of ERM. In the classical regime ($d < n$, d large but fixed), CV$_{loo}$ stability implies generalization and consistency (for $n \to \infty$). In the modern regime ($d > n$), as described in this paper, optimizing CV$_{loo}$ stability selects the minimum norm interpolating solution to the kernel least squares problem which has the best generalization performance.

The main contribution of this paper is in characterizing the stability of (possibly interpolating) solutions to the kernel least squares problem, in particular deriving excess risk bounds via a stability argument. In the process, we show that among all the interpolating solutions, the one with minimum norm also minimizes a bound on stability. Since the excess risk bounds of the minimum norm interpolating solution depend on the minimum singular value of the kernel matrix which is closely related to the condition number, we establish here a link between *numerical and statistical* stability. This also holds for solutions computed by gradient descent, since gradient descent converges to minimum norm solutions in the case of "linear" kernel methods. Our approach is simple and combines basic stability results with matrix inequalities. It is our hope that similar results may be established for deep networks, in particular with respect to minimum norm solutions being the most stable.

References

Sanjeev Arora, Simon S. Du, Wei Hu, Zhi yuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *CoRR*, abs/1901.08584, 2019.

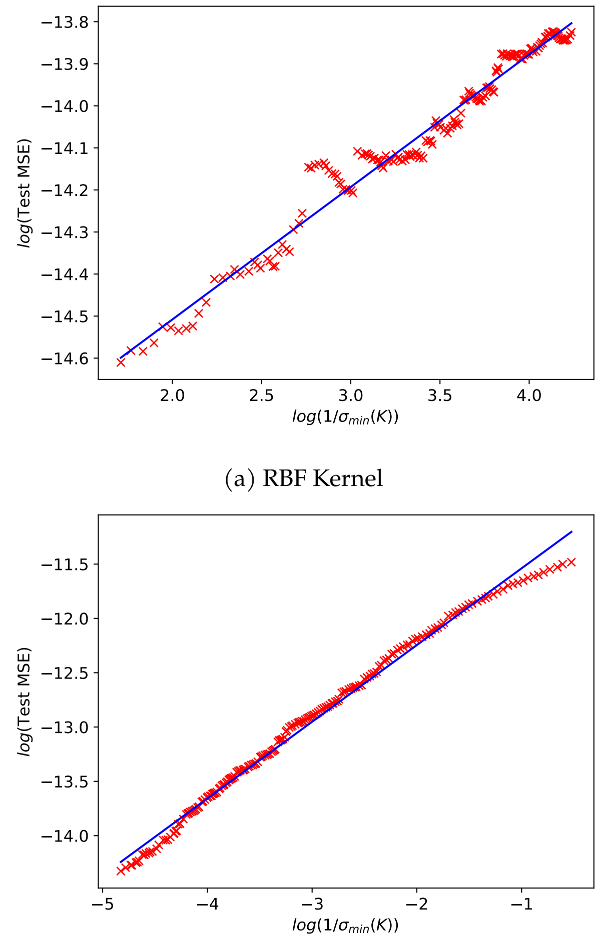
Jerzy K Baksalary, Oskar Maria Baksalary, and Götz Trenkler. A revisitation of formulae for the moore–penrose inverse of modified matrices. *Linear Algebra and Its Applications*, 372:207–224, 2003.



(a) Train and Test MSE

(b) CV$_{loo}$ Stability

Figure 1: Plot of (a) train and test mean squared error and (b) CV$_{loo}$ stability vs distance between an interpolating solution $\hat{\mathbf{w}}$ and the minimum norm interpolating solution $\mathbf{w}^\dagger$ of a linear regression problem. Data was synthetically generated as $\mathbf{y} = \mathbf{X}\mathbf{w}$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ with i.i.d $\mathcal{N}(0, 1)$ entries and $d = 1000, n = 200$. A held out test dataset of 100 samples was also generated. Other interpolating solutions were computed as $\hat{\mathbf{w}} = \mathbf{X}^\dagger \mathbf{y} + (\mathbf{I} - \mathbf{X}^\dagger \mathbf{X})\mathbf{v}$ and the norm of $\mathbf{v}$ was varied to obtain the plot. Train MSE is 0 for all interpolants, but test MSE increases as $||\mathbf{v}||$ increases, with $\mathbf{w}^\dagger$ having the best performance. CV$_{loo}$ stability also increases as $||\mathbf{v}||$ increases, with the minimum norm interpolant having the best stability. These plots represent results averaged over 100 trials.



(a) RBF Kernel

(b) Polynomial Kernel

Figure 2: Plot of log test mean squared error vs log $\frac{1}{\sigma_{\min}(K)}$ for a kernel least squares problem using (a) an RBF kernel with $\sigma = 5$ and (b) a polynomial kernel of degree 2. We synthetically generated a training dataset $\mathbf{X} \in \mathbb{R}^{n \times d}$ $(d = 20, n = 200)$ from i.i.d $\mathcal{N}(0, 1)$ entries and $y_i = \exp(-2 \left\| \mathbf{x}_i \right\|)$, as well as a held out test dataset of 100 samples. In both Figure (a) and (b), we vary the size of the training dataset from 10 to 200 to obtain kernel matrices with different singular values. These plots represent results averaged over 100 trials.

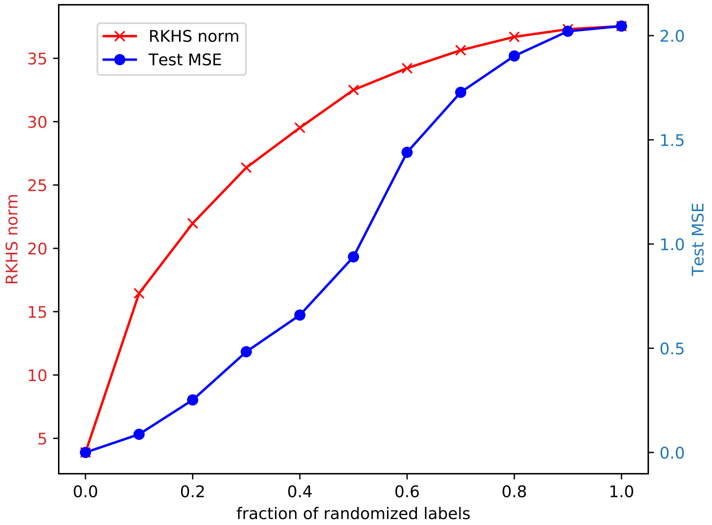


Figure 3: Effect of randomizing labels on the solution norm and test mean squared error. This plot was generated from a binary classification experiment with an RBF kernel, where the data was drawn from $\mathcal{N}(\pm 3 \times \mathbf{1}_{20}, \mathbf{I}_{20})$ with 200 training samples and 100 test samples.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *CoRR*, abs/1906.11300, 2019. URL `http://arxiv.org/abs/1906.11300`.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 541–549. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/belkin18a.html`.

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. ISSN 0027-8424. doi: 10.1073/pnas.1903070116. URL `https://www.pnas.org/content/116/32/15849`.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

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

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

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

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

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

Tengyuan Liang and Benjamin Recht. Interpolating classifiers make few mistakes. *arXiv preprint arXiv:2101.11815*, 2021.

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

Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.

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.

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

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

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

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. ISSN 1572-9044. doi: 10.1007/s10444-004-7634-z. URL `http://dx.doi.org/10.1007/s10444-004-7634-z`.

Eduard Oravkin and Patrick Rebeschini. On optimal interpolation in linear regression, 2021.

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

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

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

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

Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 28*, pp. 1630–1638. Curran Associates, Inc., 2015. URL `http://papers.nips.cc/paper/6015-learning-with-incremental-iterative-regularization.pdf`.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670, December 2010. ISSN 1532-4435.

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

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. URL `https://openreview.net/forum?id=Sy8gdB9xx`.

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 $S = \{z_1, z_2, \dots, z_n\}$, the leave one out training set $S_{-i} = \{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n\}$, and the replace one training set $(S_{-i}, z) = \{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n, z\}$

A.1 Replace one and leave one out algorithmic stability

Similar to the definition of expected CV$_{loo}$ stability in equation (3) 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) Bousquet & Elisseeff (2001). There exists $\alpha_{ro} >$ such that for all $i = 1, \dots, n$,

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

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 $i = 1, \dots, n$, S and z,

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

This observation motivates the notion of leave one out algorithmic stability (in expectation) Bousquet & Elisseeff (2001)]

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

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

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

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 14 *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 (4). ■

We can prove a similar result relating excess risk with CV$_{ro}$ stability.

Lemma 15 (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 $\mathbb{E}_S[I[f_S]] = \mathbb{E}_{S,z}[V(f_S, z)]$. ■

A.3 Discussion on Stability and Generalization

Below we discuss some more aspects of stability and its connection to other quantities in statistical learning theory.

Remark 16 (CV$_{loo}$ stability in expectation and in probability) *In Mukherjee et al. (2006), 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 Mukherjee et al. (2006), 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 (3) is what we study in the following sections.

Remark 17 (Connection to uniform stability and other notions of stability) *Uniform stability, introduced by Bousquet & Elisseeff (2001), corresponds in 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 Kutin & Niyogi (2002) , Mukherjee et al. (2006).*

Remark 18 (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 a 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 Mukherjee et al. (2006), see also Kutin & Niyogi (2002), Mukherjee et al. (2006). Indeed, a main result in Mukherjee et al. (2006) shows that CV$_{loo}$ stability is equivalent to consistency of ERM:*

Theorem 19 *Mukherjee et al. (2006) 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 20 (CV$_{loo}$ stability & in-sample/out-of-sample error) *Let $(S, z) = \{z_1, \ldots, z_n, z\}$, ($z$ is a data point drawn according to the same distribution) and the corresponding ERM solution $f_{(S,z)}$, then (4) 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 21 (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 Mukherjee et al. (2006), 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 Mukherjee et al. (2006) and is controlled in ERM as $n \to \infty$, see Theorem 19 above.

B Generalized Inverse of a Perturbed Operator

In this section we consider a linear operator perturbed by a rank-one operator, ie. $\mathbf{M} = \mathbf{A} + \mathbf{c}\mathbf{d}^\top$, where $\mathbf{M}, \mathbf{A} : U \to V$, $\mathbf{d} \in U$, $\mathbf{c} \in V$. Here U, V are inner product vector spaces over the field $\mathbb{R}$.

Theorem 22 (Theorem 6 of Meyer (1973)) *Let $\mathbf{A}^\dagger$ be the pseudoinverse of $\mathbf{A}$, and define $\mathbf{k} = \mathbf{A}^\dagger\mathbf{c}$, $\mathbf{h} = \mathbf{d}^\top\mathbf{A}^\dagger$. Let us consider the case where $\mathbf{c} \in$ range$(\mathbf{A})$, $\mathbf{d} \in$ range$(\mathbf{A}^\top)$, and $\beta = 1 + \mathbf{d}^\top\mathbf{A}^\dagger\mathbf{c} = 0$. Then the pseudoinverse $\mathbf{M}^\dagger$ of the perturbed operator is given by:*

$$\mathbf{M}^\dagger = \mathbf{A}^\dagger - \mathbf{k}\mathbf{k}^\dagger\mathbf{A}^\dagger - \mathbf{A}^\dagger\mathbf{h}^\dagger\mathbf{h} + (\mathbf{k}^\dagger\mathbf{A}^\dagger\mathbf{h}^\dagger)\mathbf{k}\mathbf{h} \tag{27}$$

Proof
We will reproduce the proof of Theorem 6 in Meyer (1973) here, with the only difference being that instead of matrices, we have linear operators.
Consider the operators $\mathbf{A}\mathbf{A}^\dagger - \mathbf{h}^\dagger\mathbf{h}$, $\mathbf{A}^\dagger\mathbf{A} - \mathbf{k}\mathbf{k}^\dagger$. These are both orthogonal projectors, since they are symmetric and idempotent. This can be checked quite easily, using the facts that $\mathbf{A}\mathbf{A}^\dagger\mathbf{h}^\dagger = \mathbf{h}^\dagger$, $\mathbf{h}\mathbf{A}\mathbf{A}^\dagger = \mathbf{h}$, $\mathbf{A}^\dagger\mathbf{A}\mathbf{k} = \mathbf{k}$, and $\mathbf{k}^\dagger\mathbf{A}^\dagger\mathbf{A} = \mathbf{k}^\dagger$. Both of these operators have their rank equal to rank$(\mathbf{A}) - 1$. This is also the case for the operator $\mathbf{M}$, from Lemma 1 of Meyer (1973).
Hence we have rank$(\mathbf{M})$ = rank$(\mathbf{A}\mathbf{A}^\dagger - \mathbf{h}^\dagger\mathbf{h})$ = rank$(\mathbf{A}^\dagger\mathbf{A} - \mathbf{k}\mathbf{k}^\dagger)$.
With the facts that $\mathbf{A}\mathbf{A}^\dagger\mathbf{c} = \mathbf{c}$, $\mathbf{h}\mathbf{c} = -1$, and $\mathbf{h}\mathbf{A} = \mathbf{d}^\top$, we have that

$$(\mathbf{A}\mathbf{A}^\dagger - \mathbf{h}^\dagger\mathbf{h})\mathbf{M} = \mathbf{M}$$

This means that range$(\mathbf{M}) \subset$ range$(\mathbf{A}\mathbf{A}^\dagger - \mathbf{h}^\dagger\mathbf{h})$.
Likewise, with the facts that $\mathbf{d}^\top\mathbf{A}^\dagger\mathbf{A} = \mathbf{d}^\top$, $\mathbf{d}^\top\mathbf{k} = -1$, and $\mathbf{A}\mathbf{k} = \mathbf{c}$, we have that

$$\mathbf{M}(\mathbf{A}^\dagger\mathbf{A} - \mathbf{k}\mathbf{k}^\dagger) = \mathbf{M}$$

and hence, range$(\mathbf{M}^\top) \subset$ range$(\mathbf{A}^\dagger\mathbf{A} - \mathbf{k}\mathbf{k}^\dagger)$. putting these together, we have that:

$$\begin{aligned}\mathbf{M}\mathbf{M}^\dagger &= \mathbf{A}\mathbf{A}^\dagger - \mathbf{h}^\dagger\mathbf{h} \\ \mathbf{M}^\dagger\mathbf{M} &= \mathbf{A}^\dagger\mathbf{A} - \mathbf{k}\mathbf{k}^\dagger\end{aligned} \tag{28}$$

If $\mathbf{X}$ is the right hand side of 27, we can use $\mathbf{h}\mathbf{A}\mathbf{A}^\dagger = \mathbf{h}$ and the above equation to obtain $\mathbf{X}\mathbf{M}\mathbf{M}^\dagger = \mathbf{X}$. We can also use the above equation, $\mathbf{k}^\dagger\mathbf{A}^\dagger\mathbf{A} = \mathbf{k}^\dagger$, $\mathbf{h}\mathbf{A} = \mathbf{d}^\top$, and $\mathbf{h}\mathbf{c} = -1$ to obtain $\mathbf{X}\mathbf{M} = \mathbf{M}^\dagger\mathbf{M}$. Since this means that $\mathbf{X}$ satisfies the two conditions of Lemma 2 in Meyer (1973), we have shown that $\mathbf{M}^\dagger = \mathbf{X}$.

■

Lemma 23 (Lemma 1 of Meyer (1973)) *Let $\mathbf{A}^\dagger$ be the pseudoinverse of $\mathbf{A}$, and define $\mathbf{u} = (\mathbf{I} - \mathbf{A}\mathbf{A}^\dagger)\mathbf{c}$, $\mathbf{v} = \mathbf{d}^\top(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})$, and $\beta = 1 + \mathbf{d}^\top\mathbf{A}^\dagger\mathbf{c} = 0$. Then the rank of the perturbed operator $\mathbf{M} = \mathbf{A} + \mathbf{c}\mathbf{d}^\top$ is given by:*

$$\textit{rank}(\mathbf{A} + \mathbf{c}\mathbf{d}^\top) = \textit{rank}\begin{bmatrix}\mathbf{A} & \mathbf{u} \\ \mathbf{v} & -\beta\end{bmatrix} - 1$$

Lemma 24 (Lemma 2 of Meyer (1973)) *If $\mathbf{X}$ and $\mathbf{M}$ are operators such that $\mathbf{X}\mathbf{M}\mathbf{M}^\dagger = \mathbf{X}$ and $\mathbf{M}^\dagger\mathbf{M} = \mathbf{X}\mathbf{M}$, then $\mathbf{X} = \mathbf{M}^\dagger$.*