CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 103 **March 26, 2020**

Stable Foundations for Learning: a foundational framework for learning theory in both the classical and modern regime.

Tomaso Poggio

Abstract

We consider here the class of supervised learning algorithms known as Empirical Risk Minimization (ERM). The classical theory by Vapnik and others characterize *universal consistency* of ERM in the *classical regime* in which the architecture of the learning network is fixed and n, the number of training examples, goes to infinity. According to the classical theory, the minimizer of the empirical risk is consistent if the hypothesis space has finite complexity. We do not have a similar general theory for the *modern regime* of interpolating regressors and overparameterized deep networks, in which $d > n$ and $\frac{d}{n}$ remains constant as n goes to infinity.

In this note I propose the outline of such a theory based on the specific notion of CV_{loo} *stability* of the learning algorithm with respect to perturbations of the training set. The theory shows that for interpolating regressors and separating classifiers (either kernel machines or deep RELU networks)

1. minimizing CV_{loo} stability minimizes the expected error
2. the most stable solutions are minimum norm solutions

The hope is that this approach may lead to a unified theory encompassing both the modern regime and the classical one.

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

Stable Foundations for Learning: a foundational framework for learning theory in both the classical and modern regime.

Tomaso Poggio

March 26, 2020

Abstract

We consider here the class of supervised learning algorithms known as Empirical Risk Minimization (ERM). The classical theory by Vapnik and others characterize *universal consistency* of ERM in the *classical regime* in which the architecture of the learning network is fixed and n, the number of training examples, goes to infinity. According to the classical theory, the minimizer of the empirical risk is consistent if the hypothesis space has finite complexity. We do not have a similar general theory for the *modern regime* of interpolating regressors and overparamerized deep networks, in which $d > n$ and $\frac{d}{n}$ remains constant as n goes to infinity.

In this note I propose the outline of such a theory based on the specific notion of CV_{loo} *stability* of the learning algorithm with respect to perturbations of the training set. The theory shows that for interpolating regressors and separating classifiers (either kernel machines or deep RELU networks)

1. minimizing CV_{loo} stability minimizes the expected error
2. the most stable solutions are minimum norm solutions

The hope is that this approach may lead to a unified theory encompassing both the modern regime and the classical one.

1 Foundations of Learning Theory

Developing theoretical foundations for learning is a key step towards understanding intelligence. Supervised learning is a paradigm in which natural or artificial networks learn a functional relationship from a set of n input-output training examples. A main challenge for the theory is to determine conditions under which a learning algorithm will be able to predict well on new inputs after training on a finite training set. What should be optimized in ERM to minimize the expected error and, for $n \to \infty$, to achieve consistency? Ideally, we would like to have theorems spelling out, for instance, that consisteny depends on constraining appropriately the hypothesis space.

Indeed 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 networks –

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.

In trying to develop a theory that can deal with the classical *and* the modern regime, it seems natural to abandon the idea of the hypotehsis space as the object of interest and focus instead on properties of the algorithms. Twenty years ago, while trying to formulate principles of learning beyond ERM (and beyond the use of measures of complexity such as VC dimension, covering numbers and Rademacher numbers), we noted [1] that any supervised learning algorithm is a map L from data sets to hypothesis functions. For a general theory, we asked: *what property must the learning map L have for good generalization error?* The answer was that LOO stability (see [1]) together with CV_{loo} stability of the algorithm, both going to zero for $n \to \infty$ is sufficient for generalization for any supervised algorithm; CV_{loo} stability alone is necessary and sufficient for generalization and consistency of ERM. At the time, the surprising connection between stability and predictivity promised a new framework for the foundations of learning theory (see also [2, 3]).

In this paper we outline how this old proposal may become a learning theory encompassing both the classical and the modern regime for ERM (extensions beyond ERM seem natural but we leave them to future work). We provide several arguments about why low expected error should correspond to stable gradient descent algorithms. In particular, an algorithm that minimizes a bound in stability should minimize the expected error if the bound is tight. Stability minimization may thus provide a unifying principle that could explain, among other properties, the predictivity of deep networks as well as the double descent curve found recently in several learning techniques including kernel machines[1].

1.1 Classical Regime

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

One of the main achievements of the classical theory was a complete characterization of the necessary and sufficient conditions for generalization of ERM, and for its *consistency* (consistency requires asymptotic convergence of the expected risk to the minimum risk achievable by functions

[1] One may argue that from the point of view of this proposal, the main role of Tikhonov regularization may be to deal with the pathological situation of $d = n$, since asymptotically the inverse of the kernel does not exist if $\lambda = 0$. Of course, presence of noise (significant SNR) has the effect of requiring regularization also for cases close to $d = n$.

in $\mathcal{H}$; for ERM, generalization is equivalent to consistency). It turns out that consistency of ERM is equivalent to a precise property of the hypothesis space: $\mathcal{H}$ has to be a *uniform Glivenko-Cantelli (uGC)* class of functions (spaces of indicator functions with finite VC dimension are a special case) of uGC .

Later work [1] showed that an apparently separate requirement – the well-posedness of ERM – is in fact equivalent to consistency of ERM. Well-posedness usually means *existence, uniqueness and stability* of the solution. The critical condition is stability of the solution. Stability is equivalent to some notion of continuity of the learning map (induced by ERM) that maps training sets into the space of solutions, eg $L : Z^n \to \mathcal{H}$. We recall the definition of *leave-one-out cross-validation (in short, CV_{loo}) stability under the distribution P_S*:

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

where $V(f, z)$ is a loss function that is Lipschitz and bounded for the range of its arguments and $z = ((x, y)$. CV_{loo} stability of an algorithm measures the difference between the errors at a point z_i when it is in the training set S of f_S wrt when is not.

It was proved [2] that *For ERM, CV_{loo} stability with β_{CV}^P and δ_{CV} in Equation 1 converging to zero for $n \to \infty$ guarantees, if valid for all P, generalization and consistency (and is in fact equivalent to them).*

Notice that CV_{loo} stability is a weaker requirement than uniform stability of Bousquet and Elissef which is sufficient but not necessary for consistency of ERM in the classical regime. Of course uniform stability implies CV_{loo} stability.

1.2 Modern Regime

Recently, a different regime has been characterized, first in neural networks [4] and then in linear and kernel regression, mainly because of the pioneering work by Belkin ([5], see also [6] and [7, 8, 5, 9, 10, 11, 12]). In this modern regime, both n (the number of training data) and d (the number of parameters) grow to infinity with $\frac{n}{d}$ constant. If $d \geq n$ there may be exact fitting of the training set and the generalization gap does not go to zero. The classical approach – based on the analysis of the hypothesis space to infer asymptotic generalization and then consistency – cannot be used because there is no fixed hypothesis space. However, the notion of stability, which refers to the algorithm and not the hypothesis space, is not affected by this problem. Since in the "classical" regime of fixed hypothesis space and $n \to \infty$, stability is important, we expect that a similar notion of stability may work in the "modern" high dimensional regime of $\frac{n}{d} < 1$.

The conjecture we discuss in this paper is that *in both cases, stability remains the key requirement for predictivity.* Maximum stability – that is minimum β_{CV}^P – is usually guaranteed during minimization of the empirical loss (that is by ERM) by complexity control under the form of regularization (possibly vanishing, as in the definition of the pseudoinverse or as implicitely provided by iterative gradient descent [13]). As we said earlier, the notion of CV_{loo} stability turns out to be necessary and sufficient for distribution independent generalization and consistency in the classical framework of ERM with a fixed hypothesis space [2, 1]. In the modern regime,

when the empirical error is zero, the definition of CV_{loo} stability seems closely related to the definition of the expected error under a specific data distribution. It is thus natural to conjecture that *minimization of stability*, in a distribution dependent way, is for ERM a sufficient condition across the classical and the modern regime for minimizing expected error. In the next section we will in fact show that CV_{loo} stability is equivalent in expectation to the expected error. Then we will discuss the conjecture that optimizing CV_{loo} stability for overparametrized networks is equivalent to selecting minimum norm solutions.

2 Stability and Expected error

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

$$\forall i \in \{1, \dots, n\} \;\; E_S \left| V(f_S, z_i) - V(f_{S^i}, z_i) \right| \leq \beta_{CV}, \tag{2}$$

where $V(f, z)$ is a loss function that is Lipschitz and bounded for the range of its arguments and $z = ((x, y)$. CV_{loo} stability of an algorithm measures the difference between the errors at a point z_i when it is in the training set S of f_S wrt when is not.

We assume here that the regressor or classifier satisfies $V(f_S, z_i) = 0$, that is they fit the training data under the appropriate loss function (e.g. square loss or classification loss, for instance the function c of [14]. Then

$$\forall i \in \{1, \dots, n\} \;\; E_S |V(f_{S^i}, z_i)| = I(f_S) \tag{3}$$

where $I(f_S)$ is the expected error of f_S.

As an example consider the case in which V is the square loss and $f_S(z_i) = W_S x_i$. Then

$$V(f_{S^i}, z_i) = (W_{S^i} x_i - y_i)^2 = (W_{S^i} x_i - W_S x_i)^2 = ((W_{S^i} - W_S) x_i)^2 \tag{4}$$

We have

Theorem 1 *(informal) For regressor (and classifiers) that achieve zero error on the training set, CV_{loo} stability in expectation is equivalent to expected error.*

3 Stability and Minimum Norm

I conjecture that the solution with the best stability among all solutions provided by ERM for the overparametrized case are minimum norm solutions. I do not know how to prove this in general. I will state it as a conjecture and support it with a few physicist-like arguments. The conjecture is

Conjecture 2 *The most stable solutions for f_S satisfying $V(f_S, z_i) = 0, \quad \forall i$ are minimum norm in the parameters.*

For later use, I recall the following result, linking minimum norm and maximum margin in the case of classification (see [15]):

Lemma 3

The maximum margin problem

$$\max_{W_K,\cdots,W_1} \min_n y_n f(W;x_n), \quad \textit{subj. to} \quad \|W_k\| = 1, \quad \forall k. \tag{5}$$

is equivalent to

$$\min_{W_k} \frac{1}{2}\|W_k\|^2, \ \textit{subj. to } y_n f(W;x_n) \geq 1, \quad \forall k, \ n = 1,\ldots,N. \tag{6}$$

3.1 Linear Regressors

The first physicist argument is for linear functions $f_S(z_i) = W_S x_i$. Fitting the training set provides the set of n equations

$$W_S X - Y = 0 \tag{7}$$

Assume $W_S \in \mathbb{R}^{1,d}$, $X \in \mathbb{R}^{d,n}$ and $Y \in \mathbb{R}^{1,n}$ with $n < d$. Then there are an infinite number of solutions for W_S given by $W_S = YX^\dagger + (I - XX^\dagger)z$ where z is any vector. The solution of minimum norm is $W_S = YX^\dagger$.

Let us show that the minimum norm solution is the most stable. The minimum norm solution among all the infinite solutions is $W_S = YX^\dagger$. In the case in which S is perturbed by deleting one data point the change ΔX in X should be small and decreasing with n. This means that $W_{S^i} = (Y + \Delta Y)(X + \Delta X)^\dagger$. Suppose X is a d, n matrix with $n < d$. Then $X^\dagger = (X^T X)^{-1} X^T$ and $(X + \Delta X)^\dagger = ((X + \Delta X)^T (X + \Delta X))^{-1}(X + \Delta X)^T$ Let us assume that $||\Delta X||$ is small and $||(X^T X)^{-1}||$ is large. Let us call $X^T X = A$, let us shorten $\Delta X = \Delta$.

Then $(X + \Delta)^\dagger \approx (A + X\Delta^T + (\Delta X^T)^{-1}(X + \Delta)^T$. Consider $(A + X\Delta)^T + \Delta X^T)^{-1} \approx A^{-1} - A^{-1}(X^T \Delta X + \Delta X^T X)A^{-1}$. Thus $(X + \Delta)^\dagger \approx [A^{-1} - A^{-1}(X^T\Delta + \Delta^T X)A^{-1}][(X + \Delta)^T]$. Putting things together and inspecting the various terms shows that $W_{S^i} = W_S + D$ where D are terms that are all contain the factor A^{-1} and delta factors in either X or Y or both. The conclusion is $||W_{S^i} - W_S|| \approx ||(XX^T) - 1(\Delta X + \Delta Y)||$. In other words stability depends on $||(XX^T)^{-1}||$ an therefore on the norm $||W||$. This proof sketch should be cleaned up to show that *the minimum norm solution is the most stable solution and viceversa.* An obvious observation is that the same argument about the behavior of $||X^\dagger||$ in [16] can be used here. It shows that for random input X, CV_{loo} stability is expected to exhibit a double-descent curve implying a double-descent curve for the expected errror.

3.2 Deep Networks

Let us first introduce some notation. We define a deep network with K layers with the usual coordinate-wise scalar activation functions $\sigma(z) : \quad \mathbf{R} \to \mathbf{R}$ as the set of functions $f(W;x) =$

$\sigma(W^K\sigma(W^{K-1}\cdots\sigma(W^1x)))$, where the input is $x \in \mathbf{R}^d$, the weights are given by the matrices W^k, one per layer, with matching dimensions. There are no bias terms: the bias is instantiated in the input layer by one of the input dimensions being a constant. We consider the case in which f takes scalar values, implying that the last layer matrix W^K is has size 1 x h_{K-1}, where h_k denotes the size of layer k. The weights of hidden layer k has size $h_k \times h_{k-1}$. In the case of of binary classification which we consider here the labels are $y \in \{-1, 1\}$. The activation function is the ReLU activation. For the network, homogeneity of the ReLU implies $f(W; x) = \prod_{k=1}^{K} \rho_k f(V_1, \cdots, V_K; x)$, where $W_k = \rho_k V_k$ with the matrix norm $||V_k||_p = 1$ and $||W_k|| = \rho_k$.

There are several physicist-like approaches to show that changes in the weights due to small changes in the training set will be proportional to the norm of the weights. A simple observation goes as follows. In a deep net, the product of the norms in a K-layer networks is $\rho_1 \cdots \rho_K$. Since we know that if the ρ_k start equal then they grow at the same rate under gradient descent and thus remain equal (see [15]), we assume that the total norm of the network is ρ^K (the argument is valid even if the ρ_k are different). Assume now that the weights of each layer are perturbed because of a change, such as leave-one-out, in the training set . Then the overall norm will change as

$$\rho^K \to K\rho^{K-1}\Delta\rho, \tag{8}$$

implying that for $V(f, z) = c_\gamma(f(x), y)$ as defined in section 4.2.2 of [17]

$$||V(f_{S^i}(x_i) - f_S(x_i))|| \leq \frac{1}{\gamma}||f_{S^i}(x_i) - f_S(x_i)||||x|| \leq \frac{1}{\gamma}\rho^{K-1}(\rho - \Delta\rho) \tag{9}$$

Thus networks with minimum norm ρ (for a fixed margin) minimize $E_S|Vf_S^i(x_i) - f_S(x_i))|$ and thus optimize CV_{loo} stability. The same argument is valid for other loss functions such as the square losss.

3.3 A General Approach?

I currently believe that a general approach to establish that stable solutions are minimum norm and viceversa may rely on the *implict function theorem* or on the more powerful *constant rank theorem*. The observation is that fitting the training set corresponds to the equation

$$F(X, Y, W) = 0 \tag{10}$$

where $X^*, Y*$ is the training set, W is the set of weights and $F(X, Y, W)$ is a set of n equations for each of the data points (columns of X and Y). Under assumptions of differentiability of F, the interpolating or separating property defines a mapping $W(X, Y)$ in the neighborhood of the solution $X^*, Y*, W*$ such that $F(X, Y, W(X, Y)) = 0$ in that neighborhood. Furthermore $\frac{\partial W}{\partial X}$ may be computed in terms of the Jacobian of F and other derivatives. This should be checked using the constant rank theorem because of possible degeneracies in the Jacobian. In the case of $F(X, Y, W) = WX - Y$, this approach would then provide $\Delta W(X) =\approx \frac{\partial W}{\partial X}\Delta X \approx X^\dagger \Delta X$. Thus

Conjecture 4 *(very informal) Using the constant rank theorem, CV_{loo} stability for kernel regressors+classifiers and for deep nets, can be bounded by the norm of the weights. Thus optimum stability is equivalent to minimum norm solutions.*

3.4 Hard margin SVM

In the case of hard margin linear SVM it is not clear in terms of the classical theory why one should select the maximum margin solution among all the separating hyperplanes. Our approach provides an answer: one must choose the most stable solutions in order to minimize the expected error, and the most stable solution is the minimum norm one for margin equal to 1 (which is equivalent to the maximum margin solution, see Lemma).

3.5 Gradient Descent and Selection of Minimum Norm Solutions

Until now we have discussed ERM, without discussing the optimization algorithm used for minimization. The summary of our results is that in order to ensure good expected error, it is necessary to select the minimum norm solutions among all the infinite solutions that achieve zero of the empirical loss. So ERM is not enough in the overparametrized case. However, it turns out that if GD is used to perform ERM, GD will select the minimum norm solution both in the case of kernel regression ([13]) and of deep networks, at least for the exponential loss (see [15]).

4 Caveats

In summary, the two main claims of this paper are 1) that minimization of stability ensures minimization of the expected erroror and 2) that minimizing stability is equivalent to choose the minimum norm solutions among all the solutions found by fitting the training set.

We now need to derive quantitative bounds for the case of kernel regressors and for deep networks. Two papers in preparation [18, 19] will describe those specific results.

5 Conclusions

In summary, optimization of CV_{loo}-type stability minimizes for $n \to \infty$ the expected error in both the classical and the modern regime of ERM. It is thus *a sufficient condition* for predictivity in ERM (but probably beyond ERM, see [1]).

In the classical regime, stability implies generalization and consistency[2]. In the modern regime, stability explains the double descent curve in kernel interpolants [18] and why maximum

[2] We think that regularization is a way to obtain an effect similar to $n > d$, especially in the asymptotically pathological case of $n = d$. We conjecture that the dynamical system associated with gradient descent algorithms has hyperbolic minima in the classical regime, including regularized ERM, but just convex minima in the modern regime, corresponding to vanishing regularization (similar to the definition of the pseudoinverse).

margin solutions in deep networks trained under exponential-type losses may minimize expected error (this does not mean they are globally optimal), see [19].

The classical conditions for consistency of ERM – such as the hypothesis space being a uGC class – can be regarded as the formalization of a 'folk theorem', which says that among simple theories the one that fits the data best should be preferred. The modern conditions would say that among all the theories that exactly fit the data, the simplest one should be preferred. The framework that unifies the classical and modern regime is stability. Given that the empirical error is good – even zero – the most stable solutions predict best. This corresponds to the statement, for both the classical and the modern regime, that stable ERM should— most of the time—change the current best model only incrementally, as new data become available.

Acknowledgments We thank Gil Kur, Andy Banbuski, Lorenzo Rosasco and especially Silvia Valle. This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216, and part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. This research was also sponsored by grants from the National Science Foundation (NSF-0640097, NSF-0827427), and AFSOR-THRL (FA8650-05-C-7262).

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

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

References

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

[2] 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.

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

[4] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016.

[5] 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.

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

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

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

[9] 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.

[10] 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.

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

[12] 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.

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

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

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

[16] T. Poggio, G. Kur, and A. Banburski. Double descent in the condition number. *CBMM memo 102*, 2019.

[17] O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In *Neural Information Processing Systems 14*, Denver, CO, 2000.

[18] Lorenzo Rosasco, Gil Kur, and Tomaso Poggio. Stability of kernel regression in the modern regime. *in preparation*, 2020.

[19] Tomaso Poggio and et al. Stability of deep networks. *in preparation*, 2020.