CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 103 **May 30, 2020**

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

Tomaso Poggio[1]

[1]Center for Brains, Minds, and Machines, MIT

Abstract

I 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. We do not have a similar general theory for the *modern regime* of interpolating regressors and overparamerized deep networks, in which $d > n$ as n goes to infinity.

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

1. minimizing cross-validation leave-one-out stability minimizes the expected error
2. minimum norm solutions are the most stable solutions

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

This material is based upon work 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

May 30, 2020

Abstract

I 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. We do not have a similar general theory for the *modern regime* of interpolating regressors and overparamerized deep networks, in which $d > n$ 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 suggests that for interpolating regressors and separating classifiers (either kernel machines or deep RELU networks)

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

The hope is that this approach may lead to a 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 I sketch 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 I leave them to future work). I provide several arguments about why low expected error should correspond to stable gradient descent algorithms in the moder n regime. In particular, an interpolating algorithm that minimizes a bound on CV_{loo} stability should minimize the expected error. Stability optimization 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 in $\mathcal{H}$; for ERM, generalization is equivalent to consistency). It turns out that consistency

[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 $d = n$.

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 .

Our 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.

We 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 the 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, I expect that a similar notion of stability may work in the "modern" high dimensional regime of $\frac{n}{d} < 1$.

The conjecture discussed 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 I 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 for interpolating algorithms (under specific data distributions). 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 I will show that CV$_{loo}$ stability is almost equivalent in expectation to the expected error for interpolating regressors or classifiers. Then I will discuss the separate conjecture that optimizing CV$_{loo}$ stability for overparametrized networks is equivalent to selecting minimum norm solutions.

2 Stability and Expected error

Let us recall the definition *in expectation of leave-one-out cross-validation (in short, CV$_{loo}$) stability under the distribution P_S*:

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

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 want now to consider the case – typical for overparametrized models – of interpolating regressors or separating classifiers, that is the case in which the regressors or classifiers can usually satisfiy $V(f_S, z_i) = 0$, 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]). The idea is that then the first term in Equation 2 is negligeable for specific distributions of the data and CV$_{loo}$ stability becomes essentially equal, in expectation, to the expected loss. This intuition, however, needs to be made rigorous.

To do so, I use the following positivity property of exact ERM [2]

$$V(f_{S^i}, z_i) - V(f_S, z_i) \geq 0. \tag{3}$$

Then $E_S \, |V(f_S, z_i) - V(f_{S^i}, z_i)| = E_S[V(f_{S^i}, z_i)] - E_S[V(f_S, z_i)]$

Thus

$$\forall i \in \{1, \dots, n\} \;\; E_S \, |V(f_S, z_i) - V(f_{S^i}, z_i)| = E_S I[f_{S^i}] - E_S I_S[f_S] \tag{4}$$

where $I(f_S)$ is the expected error of f_{S^i} and $I_S[f_S]$ is the empirical error of f_S. Under specific assumpions on the algoritm and the distribution P_S, the term $E_S I_S[f_S]$ can be negligeable, as we will see later. In these cases, CV$_{loo}$ stability is indeed equal to $I[f_{S^i}]$. In turn for ERM $I[f_{S^i}]$ converges in probability to $I[f_S]$ for $n \to$
$infty$. As an example of the $I[f_{S^i}]$ term, consider the case in which V is the square loss and $f_{S^i}(z_i) = W_{S^i} x_i$. Then

$$V(f_{S^i}, z_i) = (W_{S^i} x_i - y_i)^2 = (W_{S^i} x_i - W_S x_i)^2 = ((W_{S^i} - W_S) x_i)^2 \tag{5}$$

We have

Theorem 1 *(very informal) For distributions P_S for which a given regressor (or classifier) has an expected zero empirical error on the training set, CV_{loo} stability in expectation is equivalent to expected error of the regressor (or classifier).*

Remark

The same result can be obtained for *quasi-ERM*, which selects an almost minimizer of the empirical risk, in the limit of $n \to \infty$ by using the *almost positivity* property of quasi-ERM.

In the following we consider the expected error term in CV$_{loo}$ stability, effectively assuming that the empirical error is negligeable.

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 full generality. I will state it as a conjecture and support it with a few specific cases. 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

$$\max_{W_K,\cdots,W_1} \min_n y_n f(W; x_n), \quad \textit{subj. to} \quad \|W_k\| = 1, \quad \forall k. \tag{6}$$

is equivalent to

$$\min_{W_k} \frac{1}{2}\|W_k\|^2, \ \textit{subj. to} \ y_n f(W; x_n) \geq 1, \quad \forall k, \ n = 1, \ldots, N. \tag{7}$$

3.1 Linear Regressors

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

$$W_S X - Y = 0 \tag{8}$$

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 explain the intuition 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 $X^T X = A$ and $\Delta X = \Delta$.

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 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 $f(W; x) = \sigma(W^K \sigma(W^{K-1} \cdots \sigma(W^1 x)))$, 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 ways 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

$$\rho^K \to K\rho^{K-1}\Delta\rho, \tag{9}$$

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

$$V(f_{S^i}(x_i) - f_S(x_i)) \leq \frac{1}{\gamma} ||f_{S^i}(x_i) - f_S(x_i)|| ||x|| \leq \frac{1}{\gamma}\rho^{K-1}(\rho - \Delta\rho) \tag{10}$$

Thus minimizing the norm ρ (for a fixed margin) minimizes a bound on $E_S|Vf_S^i(x_i) - f_S(x_i))|$, that is on CV$_{loo}$ stability. The same argument is valid for other loss functions such as the square losss.

3.3 A General Approach?

A possibly more 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

$$F(X, Y, W) = 0 \tag{11}$$

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 (there are separate arguments, such as the perceptron learning theorem) 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 in the case of hard margin linear SVM is the minimum norm one for margin equal to 1 (which is equivalent to the maximum margin solution, see section in [15] on maximum margin and minimum norm).

3.5 Gradient Descent (GD) and Selection of Minimum Norm Solutions

Until now I have discussed ERM, without discussing the optimization algorithm used for minimization. The summary is that in order to ensure good expected error for interpolating regressors, it is necessary to select the most stable solution and to do that one should select the minimum norm solution among all the infinite solutions that achieve zero the empirical loss. So ERM is not enough by itself in the overparametrized case. However, it turns out that if GD is used to perform ERM, GD will select among the empirical minimizers the one with minimum norm both in the case of kernel regression ([13]) and of deep networks, if the loss is of the exponential type (see [15]).

4 Caveats

In summary, the two main claims of this paper are 1) that minimizing CV$_{loo}$ stability minimizes the expected erroror and 2) that minimizing stability corresponds to choosing the minimum norm solutions among all the solutions with zero empirical error.

It is now important to derive more formal bounds for both the case of kernel regressors and the case of deep networks. Two papers in preparation will [18, 19] describe those 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. In the modern regime, stability probably explains the double descent curve in kernel interpolants [18] and why maximum margin solutions in deep networks trained under exponential-type losses may minimize expected error (this does not mean they are globally optimal), see [19].

Conditions for learnability and stability in learning theory may have deep, almost philosophical, implications: as remarked by V. Vapnik, they can be regarded as equivalent conditions that guarantee any scientific theory to be predictive and therefore "scientific". The condition coming from classical learning theory corresponds to choosing the theory from a fixed “small” set of theories that best fit the data. The condition prescribed by the modern theory corresponds to choosing a theory from a “large” hypothesis set (that can even increase before new data arrive) that fits the data *and* is simplest (Occam razor, Einstein). These two conditions can be summarized and unified by the principle of selecting the most stable theory — the one that most of the time changes the least if data are perturbed or when new data arrive. Thus Thomas Kuhn scientific revolutions are allowed, as long as they do not happen too often!

Acknowledgments We thank Gil Kur, Andy Banbuski, and especially Silvia Villa and Lorenzo Rosasco. 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.

[20] Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes. Fisher-rao metric, geometry, and complexity of neural networks. *CoRR*, abs/1711.01530, 2017.

[21] Yaim Cooper. The loss landscape of overparameterized neural networks. *CoRR*, abs/1804.10200, 2018.

[22] Lei Wu, Zhanxing Zhu, and Weinan E. Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes. *arXiv e-prints*, page arXiv:1706.10239, June 2017.

[23] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *arXiv e-prints*, page arXiv:1609.04836, September 2016.

[24] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects. *arXiv e-prints*, page arXiv:1803.00195, February 2018.

[25] Y. Cooper. The critical locus of overparameterized neural networks. *arXiv e-prints*, page arXiv:2005.04210, May 2020.

6 Appendix: SGD and deep overparametrized nets

This is just to date some interesting ideas for another related paper while hiding them for a little while. They are about optimization of deep networks using SGD.

6.1 Background: gradient descent and SGD

Consider

$$\min_W L(f(W)) = \min_W \frac{1}{N} \sum_n^N \ell_n \tag{12}$$

with $\ell_i = (y_i - f(W; x_i))^2$.

GD can be used to minimize $L(f(W))$ by running the following dynamical system (e.g. gradient flow)

$$\dot{W} = \nabla_W L(f(W)) = \sum_n^N \nabla_W f(W; x_n)(y_n - f(W; x_n)). \tag{13}$$

SGD can be formulated as follows. First define

Definition 5 *A random vector $v \in R^d$ drawn from a distribution $\mathcal{D}$ is a sampling vector if* $\mathbb{E}_{\mathcal{D}}[v_i] = 1 \quad \forall i$

Then the stochastic version of Equation 12 is

$$\min_W \mathbb{E}_{\mathcal{D}}[L(f(W))] = \min_W \mathbb{E}_{\mathcal{D}} \sum_n^N v_n \ell_n \tag{14}$$

Usually the distribution over $\mathcal{D}$ is assumed to be random v with independent components v_i, satisfying condition 5. This implies that in expectation SGD is equal to GD.

6.2 Critical points

Finding the interpolating global minimizers of $L = \sum \ell_i$ is equivalent to finding the set of network weights W^* that solve the system of equations $\ell_i(W^*) = 0 \quad \forall i = 1, \cdots, N$. Thus instead of finding all the critical points of the gradient of L, we would like to find the joint minimizers – that is the W – that minimize $\ell_i \quad \forall i = 1, \cdots, N$.

We distinguish two sets of solution to $\nabla L = 0$:

1. solutions of $\nabla \ell_i = 0, \forall i$
2. solutions of $\sum_n (y_n - f(x_n)) \nabla f(x_n) = 0$ that are not solutions of $\nabla \ell_i = 0, \forall i$

The solutions 1) of $\nabla \ell_i = ((f(x_i) - y_i)\nabla_W f(x_i) = 0, \forall i$ consist of the global minima that is $((f(x_i) - y_i) = 0, \forall i$ and of other points, which we call here "spurious" critical points, for which $\nabla \ell_i = 0, \forall i$ but $L \neq 0$. This can happen if $\nabla_W f(x_i) = 0$ for some i and $(f(x_i) - y_i = 0$ for the other i. These solutions have interesting property that *if* $\nabla_W f(x_i) = 0$ *for some* i *then* $f(x_i) = 0$ *for the same* i.

The proof of this fact uses the structural lemma (Lemma 2.1 of [20], closely related to Euler's theorem for homogeneous functions)

$$\sum_{i,j} W_k^{i,j} \left(\frac{\partial f(W;x)}{\partial W_k^{i,j}} \right) = f(W;x) \tag{15}$$

where W_k is here the vectorized representation of the weight matrices W_k for layer k. Setting $\nabla_W f(x_i) = 0$ in Equation 15 gives

$$0 = f(W;x_i) \tag{16}$$

.

6.3 Why SGD does not get stuck in critical points which are not global minimizers

The global minima and the critical points satisfying $\nabla_w \ell_i = 0, \quad \forall i$ are critical points of the loss for any subset of the training points, that is for any of the batches used in SGD. This is *not true* for the solutions 2) of $\sum_n (y_n - f(x_n))\nabla f(x_n) = 0$: they are not (generically) critical points of any random subset of the training points. This means that SGD will never stop after it reaches them (generically).

Consider

$$\dot{W} = \sum_n^N (y_n - f(x_n))\nabla f(x_n) \tag{17}$$

where $v_n(t)$ is the n-th component of v which is a random binary vector with $(N - N_{SGD})$ zero's and N_{SGD} one's, where N_{SGD} is the size of the minibatches used by SGD. For simplicity, assume the basic form of SGD, in which the minibatch size is 1. The random vector v (all zero with a single component equal to one) changes at every interation. Suppose now that $W(t)$ has reached a critical point of the gradient which is a global minimum. Then at $t+1$, $\dot{W} = 0$ independently of the choice of $v(t+1)$ since $\nabla f(x_n)$ is zero for each of the data points. Suppose instead that $W(t)$ has reached a critical point of the gradient which is not a global minimum (and is not a spurious critical point). Then at $t+1$, $\dot{W} \neq 0$ for a random choice of $v(t+1)$; if $\dot{W} = 0$ for that particular choice, in one of the subsequent iterations the random v will yield $\dot{W} \neq 0$ and the dynamical system will move out of the critical point.

Theorem 6 *(informal) Consider a noiseless situation.The dynamical systems defined by SGD and GD stops at global minima. GD will also get stuck at all other critical points of the gradient. SGD will not get stuck at other critical points apart from the spurious ones.*

6.4 Degeneracy of global minima

We are interested in finding the global minimizers achieving zero loss of

$$L(f(W)) = \frac{1}{N}\sum_{n}^{N} \ell_n \tag{18}$$

with $\ell_i = (y_i - f(W; x_i))^2$. The network f is assumed to be overparametrized with a number of weights $D >> N$ and to be able to interpolate the training data achieving $L(f(W^*)) = 0$ which implies $\ell_i = 0 \quad \forall i = 1, \cdots, N$.

If we assume overparametrized networks with $d >> N$, where d is the number of parameters and N is the number of data points, [21] proved that the global minima of $L(w)$ are highly degenerate with dimension $d - N$[2].

Theorem 7 *([21]) For an overparametrized f with smooth activation functionassuming a square loss, the minimizers $W*$ are highly degenerate with dimension $D - N$.*

This results may explain some puzzling findings:

- SGD tends to select so-called flat minima [22];
- small-batch SGD methods consistently converge to flat minimizers [23];
- most interesting is that minima found by GD can be unstable for SGD [24]. Often switching from GD to SGD at a point close to a global minimum, SGD escapes from that minimum and converges to a better minimum [22]

6.4.1 Degeneracy of global critical points

The question is whether the solutions of the global critical points, e.g. $\sum_n (y_n - f(x_n))\nabla \ell_n = 0$ that are neither global zeros nor solutions of $\nabla \ell_i = 0, \quad \forall i$ are degenerate. The answer is provided by [25] with estimates of the dimensionality of the critical points. For networks with a sufficient number of layers and sufficiently overparametrized the degeneracy of the critical points which are not global minima is less than the degeneracy of the global minima.

6.4.2 Characterization of the "spurious" critical points $\nabla \ell_i = 0, \forall i$ with $L \neq 0$

On the other hand, the degeneracy of each of the spurious critical points should be the same as the degeneracy of the global minima (in principle, there are many more spurious critical points because they are all the combinations of $f(x_i) = 0$ for some i and $\ell_i = 0$ for the remaining is).

[2] This result is also what one expects from Bezout theorem for a deep polynomial network. As Terry Tao says in his blog "from the general "soft" theory of algebraic geometry, we know that the algebraic set V is a union of finitely many algebraic varieties, each of dimension at least d-n, with none of these components contained in any other. In particular, in the underdetermined case n<d , there are no zero-dimensional components of V , and thus V is either empty or infinite".

Most of the spurious critical points are not very close to the global minimum because for each of the points for which $f(x_i) = 0$ the loss is $\ell_i = y_i^2$.

As critical points of L one can ask: which kind are they? The answer is that *they are local minima of the loss*, apart from the critical points $\nabla f(x_i) = 0, \forall i$, which is a saddle with indefinite Hessian. The argument is as follows. For some i $f(x_i) = y_i$ and $f(x_i) = 0$ for the other i; this means that $L \neq 0$. Are these points local minima or saddles? For the i for which $f(x_i) = 0$ $H(x_i) = 0$; for the other i H is positive semidefinite. Thus the Hessian associated with L, since L is the sum of the ℓ_i, is positive semidefinite. This means that the associated critical points are *degenerate local minima*.

6.4.3 The peculiar case of exponential-type loss functions

Consider the exponential loss $L = \frac{1}{N}\sum_n^N \ell_n$ with $\ell_n = e^{y_i \rho f_V(x_i)}$. The critical points of L are

$$\dot{W}_k = -\nabla_{W_k} L = \sum_n^N y_n \nabla_{W_k} f(x_n) e^{-y_n f(x_n)}. \tag{19}$$

For binary classification $y_n = \pm 1$ and f is either > 0 or < 0. In general for separating solutions, L is never $= 0$ at finite times: the dynamical system converges towards a zero infimum of L for $\rho \to \infty$ which corresponds to $t \to \infty$. However, $\dot{W}_k = 0$ if $f(x_n) = 0, \forall n$. These critical points are local minima of the loss and critical points for both GD and SGD. Unlike the square loss case, they only happen if $f(x_{)} = 0$ for all x_i, which means that the network is "dead" for any of the inputs x_i. Initializations with small, random Ws try to avoid this situation which is made even more unlikely by the sotmax operation which selects between $f < 0$ and $f > 0$ braking the symmetry in the case $f = 0$. It seems that spurious minima are much less of a problem for the exponential loss than for the square loss. The situation for cross entropy and for the logistic loss (cross entropy in the binary case) is similar to the exponential loss case since for the logistic $\ell(x) = \log(1 + e^{-yf(x)}$ with $y = \pm 1$ and $-\nabla_W \ell(x) = \frac{e^{yf}}{1+e^{yf}} \nabla_W f$.