CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 112 August 17, 2020

Implicit dynamic regularization in deep networks

Tomaso Poggio, Qianli Liao

Abstract

Square loss has been observed to perform well in classification tasks, at least as well as crossentropy. However, a theoretical justification is lacking. Here we develop a theoretical analysis for the square loss that also complements the existing asymptotic analysis for the exponential loss.

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

Implicit dynamic regularization in deep networks

Tomaso Poggio, Qianli Liao

Abstract

Square loss has been observed to perform well in classification tasks, at least as well as crossentropy [1]. However, a theoretical justification is lacking. Here we propose an analysis for gradient descent under the square loss in RELU networks that complements the existing asymptotic analysis for exponential-type loss functions [2]. In particular we predict

- that there exist an initial transient phase (TP) of regularization resulting from the nonlinear dynamics of the norm of the layers in RELU networks;
- that the TP lasts longer for smaller initializations and for deeper networks keeping the norm small;
- that a second, more classical, implicit regularization follows, ensuring convergence to a local minimum norm solution.

1 Introduction

It seems that any explanation of the ability of deep networks to be predictive requires the identification of a hidden mechanism of complexity control at work during the training of deep networks, ensuring CV stability of the solution (good stability implies good test error[3]).

In the case of exponential-type loss functions such a mechanism has been identified in the margin maximization effect of minimizing exponential-type loss functions [4, 5, 6]. However, this mechanism cannot explain the good empirical results that can be obtained using the square loss [1], assuming of course that specific forms of gradient descent used in such experiments – such as momentum or batch normalization – are not hiding complexity control effects.

In trying to solve this puzzle, we identify an interesting dynamics constraining the norm of the network that strictly depends on the presence of multiple layers and has properties matching several empirical observations. We conjecture that this transient phase of norm control (together with the implicit regularization of iterative gradient descent once close to a degenerate minimum) may have a significant role in explaining the stability of trained deep networks.

This note is a preliminary communication with more details, proofs and experiments to follow in a future update.

2 The dynamics of the norm and of the normalized weights

2.1 Gradient descent

The natural approach to training deep networks for binary classification using the square loss is to use stochastic gradient descent to find the weights W_k that minimize $L = \frac{1}{N}\sum_n \ell_n^2 = \frac{1}{N}\sum_n^N (g(x_n) - y_n)^2$, with $y = \pm 1$. In this note, we consider gradient descent instead of stochastic gradient descent. We also assume NOT to use batch normalization, momentum, weight decay, data augmentation (although the dynamic for the weight normalization algorithm is very similar).

Gradient descent on $\mathcal{L} = \frac{1}{N}(\sum_n g_n^2 - 2\sum_n y_n g_n + N)$ gives

$$\dot{W}_k = -\frac{\partial L}{\partial W_k} = -\frac{2}{N}\sum_n g_n \frac{\partial g_n}{\partial W_k} + \frac{2}{N}\sum_n y_n \frac{\partial g_n}{\partial W_k} \tag{1}$$

We now consider separately the dynamics of the norms and of the normalized weights.

2.2 Notation and assumptions

- We set $g(x) = \rho g_V(x)$ with ρ, V, g_V defined as in [6];
- in the following we use the notation f_n meaning $g_V(x_n)$, that is the normalized network;
- we assume $||x|| = 1$ implying $||g_V(x)|| = ||f(x)|| \leq 1$ at convergence;
- we assume that at initialization all the layers have the same norm, that is ρ_k is the same for all k at initialization.

2.3 Dynamics of norm and normalized weights

We consider the dynamical system induced by GD on a deep net with RELUs. We change variables by using $W_k = \rho_k V_k$, $||V_k|| = 1$. Following the calculations in [6], the following identies hold: $\frac{\partial \rho_k}{\partial W_k} = V_k^T$ and $\frac{\partial g_n}{\partial W_k} = \frac{\rho}{\rho_k}\frac{\partial f_n}{\partial V_k}$. This implies

$$\dot{W}_k = -\frac{2}{N}\frac{\rho}{\rho_k}[\sum_n (\rho f_n - y_n)\frac{\partial f_n}{\partial V_k}] \tag{2}$$

Thus gradient descent on $L = \mathcal{L} = \sum_n (\rho f_n - y_n)^2$ with the constraints on V_k and ρ_k yields the dynamical system (with $\dot{W}_k = -\frac{\partial L}{\partial W_k}$)

$$\dot{\rho_k} = \frac{\partial \rho_k}{\partial W_k}\dot{W}_k = V_k^T \dot{W}_k = -2\sum_n (\rho_k^L f_n - y_n) f_n \rho_k^{L-1} = -2\rho_k^{L-1}[\sum_n \rho_k^L (f_n)^2 - \sum_n f_n y_n] \tag{3}$$

and, with $S_k = I - V_k V_k^T$,

$$\dot{V_k} = \frac{\partial V_k}{\partial W_k}\dot{W}_k = \frac{S_k}{\rho_k}\dot{W}_k = -2\frac{\rho}{\rho_k^2}\sum_n (\rho f_n - y_n)(\frac{\partial f_n}{\partial V_k} - V_k f_n). \tag{4}$$

Because of the third assumption we can use the following

Lemma 1 $\frac{\partial \rho_k^2}{\partial t}$ *is independent of* k*.*

to claim that all ρ_k are the same at all times. Thus $\rho = \rho_k^L$, where L is the number of layers.

We use Equation 3 to derive the dynamics of $\rho = \rho_k^L$ in terms of $\dot{\rho} = \sum_k \frac{\partial \rho}{\partial \rho_k} \dot{\rho}_k$

Thus

$$\dot{\rho} = 2L\rho^{\frac{2L-2}{L}}[-\sum_n \rho (f_n)^2 + \sum_n f_n y_n] \tag{5}$$

which has a potentially very interesting dynamics as shown in the simplified simulations of Figure 1. The equilibrium value for $\dot{\rho_k} = 0$ is

$$\rho_{eq} = \frac{\sum_n y_n f_n}{\sum_n f_n^2}. \tag{6}$$

Similarly for V_k:

$$\dot{V_k} = -2\rho^{\frac{L-2}{L}} \sum_n (\rho f_n - y_n)(\frac{\partial f_n}{\partial V_k} - V_k f_n) \tag{7}$$

At equilibrium for V_k – that is when $\dot{V_k} = 0$ – the equation gives (with $\ell_n = \rho f_n - y_n$ and assuming $\sum f_n \ell_n \neq 0$)

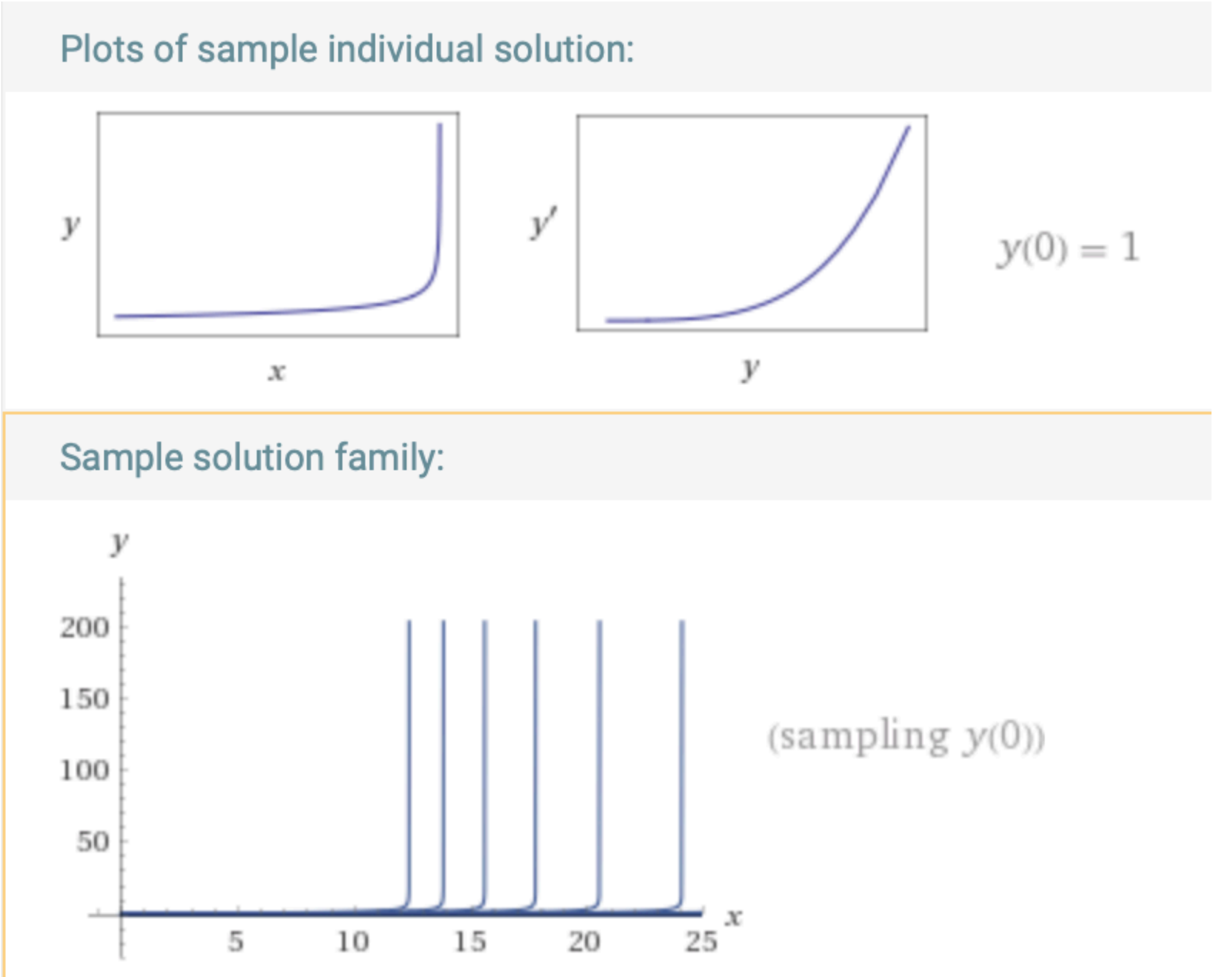
$$\sum_n (\rho f_n - y_n)(\frac{\partial f_n}{\partial V_k}) = \sum_n (\rho f_n - y_n)(V_k f_n) \tag{8}$$

Since $\sum_n (\rho_{eq} f_n - y_n) f_n = 0$ from Equation 6, it follows that at $\dot{V_k} = 0$ we obtain $\sum_n \ell_n \frac{\partial f_n}{\partial V_k} = 0$. This condition is consistent with equilibrium at ρ_{eq} since it implies, by multipying it from the left with V^T, $\sum_n \ell_n f_n = 0$.

3 Conjectures and predictions

To give an intuition of what we expect, given the previous analysis, let us assume that the initial conditions are $\rho_{t=0} \approx 0$ (but $\rho_{t=0} > 0$) and at least some of the $y_n f_n < max_V y_n f_n = 1$. Then $\rho(t)$ eventually grows (most of the time, when it soes not go to zero), but very slowly for a longish time until it grows very quickly. The dynamics of Equation 5 is that the smaller $\rho_{t=0}$ is, the longer it takes to ρ to grow (this phenomenon increase with larger L). Thus ρ is constrained by the nonlinear dynamics to be very small for a transient phase T of SGD iterations (as we mentioned, T is longer with more layers and longer with smaller initialization). During this time gradient flow on W_k is trying to minimize $\sum_n (\rho f_n - y_n)^2$ under the norm constraint. At around T, ρ will grow quickly to a value which depends on $\frac{\sum y_i f(x_i)}{\sum f^2(x_i)}$. The intuition is that this dynamics (from $t = 0$ to $t = T$) is similar to minimizing the square loss under the constraint of a small ρ that is

$$\min_{W_k} L = \sum (\rho f(x_i) - y_i)^2 \quad \text{s.t.} \rho \leq C \tag{9}$$



Plots of sample individual solution:

$y(0) = 1$

Sample solution family:

(sampling $y(0)$)

Figure 1: *Plot of y which is the square root of ρ for $L = 4$, assuming that $\sum f_n^2$ and $\sum y_n f_n$ are constant; the bottom plots show the effect of different initial conditions.*

2.pdf

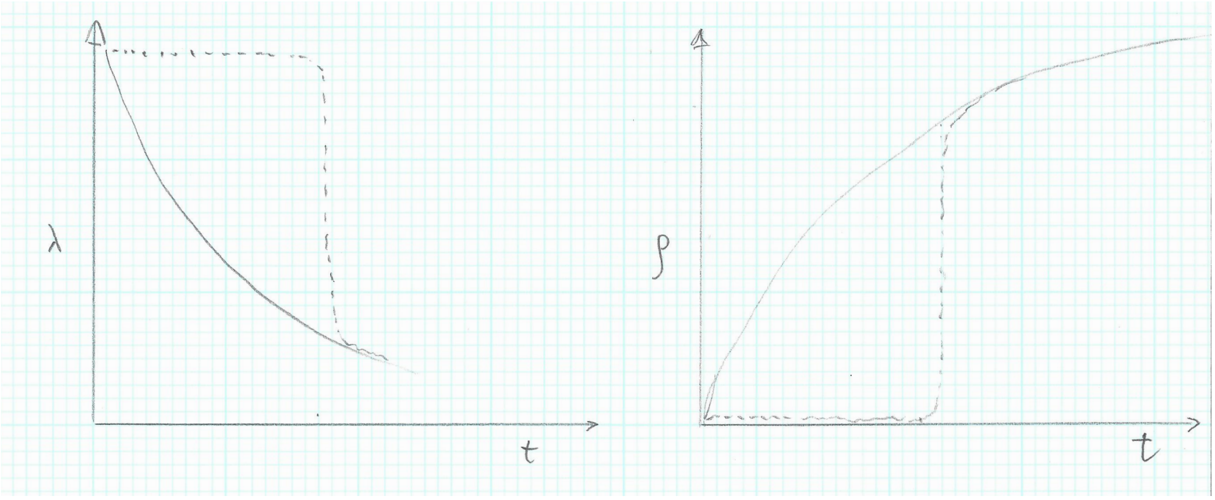


Figure 2: *On the left: plot of λ in a Tikhonov regularization hypothetically equivalent to the nonlinear dynamics of Equation 5. On the right is the corresponding dynamics for ρ.*

which is Ivanov regularization, which is itself "equivalent" to Tikhonov regularization with appropriate $\lambda(\rho)$ (see Figure 2).

The main difference is that in our case the constraint is only applied for a time T. What happens after T? To answer this question, let me recall the behavior of iterative gradient descent in the case of linear networks (see for instance [7]). In this case, the behavior of gradient descent is equivalent to a vanishing regularization with $\lambda \approx \frac{1}{T}$ where T is the number of iterations; convergence is to the minimum norm solution, *if* initialization is around zero norm. In the nonlinear case, the same behavior could be expected once GD is close to a minimum because then the loss should be locally equivalent to a Morse-Bott function (we are thinking about a positive definite Hessian in some directions and degenerate in the others).

The problem is that, unlike the linear case, the weights associated with the degenerate directions at the minimum may have grown to non-zero values during the iterations before reaching a neighborhood of the minimum. A constraint of small norm as in equation 9 for the initial T iteration may be key in ensuring a small norm before reaching a minimum: iterative gradient descent will then ensure convergence to a local linear pseudoinverse (defined in terms of the Jacobians for the nondegenerate directions).

3.1 Remarks

- Notice that $\dot{\rho_k} > 0$ as long as ρ is not much larger than 1 and $\sum_n f_n y_n > 0$.
- The lowest value of ρ_k at equilibrium ($\dot{\rho_k} = 0$) is $\rho_k = 1$ which can be achieved if $y_n f_n$ is either $= 1$ or $= 0$.

- Values $y_n f_n = 1, \rho = 1$ are stationary points of the dynamics of V_k given by $\dot{V}_k = 0$: they are minimizers with zero square loss. Notice that, in general, classification with maximum margin is not minimum norm interpolation of the labels.

- The dynamics of the "weight normalization" gradient descent algorithm can be derived using Lagrange multipliers: it has the same dynamics as described here and a slightly different dynamics for V_k (see [6]).

- Equation 6 is a critical point for the dynamics of ρ under GD but *NOT* under SGD. In the case of SGD the asymptotic value of ρ for fixed $\sum f_i y_i$ c an be expected to fluctuate randomly around the $\frac{\sum_n y_n f_n}{\sum_n f_n^2}$. Similar comments also apply to the dynamics of V_k.

3.2 Open questions

- For GD under the exponential loss, in addition to the same initial dynamic regularization, we expect a margin maximization effect at long times as shown in [2]. Thus deep nets under the square loss are more likely to overfit at long times than under exponential-type loss functions (unless momentum or regularization is used). As a consequence, early stopping is more likely to be effective for the square loss than for exponential-type loss functions.

- Does the TP regularization we have identified play a major role in obtaining solutions with good predictive performance? If yes, this may also explain why recurrent networks work almost as well as unrolled networks: the dynamic is effectively the same, though the number of parameters is much higher in the second case.

- Equation 5 implies an especially interesting dynamics for ResNets, since they are equivalent to the combination of a set of neworks with different depths from 2 to L layers and shared weights. Does this provide a more gradual control of the norm?

- Separability ($y_n f_n > 0, \forall n$) and small ρ ($\rho > 1$ but close to 1) may imply a bias towards large and small values of f_n across n (remember that $||x||_2 \leq ||x||_1 \leq d^{\frac{1}{2}}||x||_2$). The question is why the dynamics should be biased towards $\min \rho$ provided that $\rho f_i \geq 1, \quad \forall i$.

- Suppose we control ρ_k independently of V_k and of equation 5: will this lead to solutions with better generalization?

Acknowledgments This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216, and part by C-BRIC, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. This research was also sponsored by grants from the National Science Foundation (NSF-0640097, NSF-0827427), and AFSOR-THRL (FA8650-05-C-7262) and was also supported by the Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00079, Department of Artificial Intelligence(Korea University)).

References

[1] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv e-prints*, page arXiv:2005.08054, May 2020.

[2] Tomaso Poggio, Andrzej Banburski, and Qianli Liao. Theoretical issues in deep networks. *PNAS*, 2020.

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

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

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

[6] 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.

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