Implicit dynamic regularization in deep networks

Tomaso Poggio, Qianli Liao, Lorenzo Rosasco (?)

Abstract

Square loss has been observed to perform well in classification tasks, at least as well as cross-entropy for which an asymptotic analysis is available. However, a theoretical justification is lacking. Here we discuss several observations on the dynamics of gradient descent under the square loss in RELU networks. We show how convergence to a local minimum norm solution is expected when normalization techniques such as Batch Normalization (BN) or Weight Normalization (WN) are used, in a way which is similar to the behavior of linear degenerate networks under GD, though the reason for zero-initial conditions is different.
Implicit dynamic regularization in deep networks

Tomaso Poggio, Qianli Liao, Lorenzo Rosasco(?)

Abstract

Square loss has been observed to perform well in classification tasks, at least as well as crossentropy \[1\] for which an asymptotic analysis is available (see \[2\] and references therein). However, a theoretical justification is lacking. Here we discuss several observations on the dynamics of gradient descent under the square loss in RELU networks. We show how convergence to a local minimum norm solution is expected when normalization techniques such as Batch Normalization (BN) or Weight Normalization (WN) are used, in a way which is similar to the behavior of linear degenerate networks under GD, though the reason for zero-initial conditions is different. The property of the minimizer that bounds its expected error is its norm and the rank of the local Jacobian, because among all the interpolating solutions the ones associated with small Frobenius norms of the weight matrices have better margin and better expected classification error.

1 Introduction

We start from the assumption that an explanation of the ability of deep RELU networks to be predictive requires the identification of a hidden mechanism of complexity control at work during the training of deep networks\[1\].

In the case of exponential-type loss functions such a mechanism has been identified in the asymptotic 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 \[1\]\[2\];
• cannot explain non-asymptotic convergence for cross-entropy loss minimization and its dependence on initialization.

Thus we focus here on the square loss. We will discuss the dynamics of weight normalization, since mechanisms such as BN and WN seem essential for reliably training deep networks\[7\].

We conjecture here that the explanation of implicit regularization for overparametrized deep nonlinear networks is similar to why GD converges to a minimal norm solution in the case of

---

1 Complexity control should ensure CV stability of the solution (good stability implies good test error\[3\]).
2 Assuming that specific forms of gradient descent used in such experiments – such as momentum or batch normalization – are not hiding complexity control effects.
overparametrized linear systems. Crucially, our explanation requires commonly used GD-based optimization algorithms such as BN or WN. We know that for overparametrized linear systems GD converges to the minimum norm solution for zero initial conditions but also that in the presence of normalization of the weights – by BN or similar algorithms – the convergence is independent of initial conditions (see Appendix). Thus it is natural to conjecture that this same mechanism ensures generalization in a nonlinear system which can be linearized around a (degenerate) minimum.

We have also to explain why implicitly regularized regression works well for classification. Why and under which conditions there is a bias for larger margin solutions? We propose that the bias for large margin interpolating solutions that minimize the square loss depends on small initialization. The observation is that the norm $\rho$ of the solution (product of the Frobenius norms of the weight matrices of the network layers) is inversely related to the average margin. Under small initialization, GD will explore global minima with $\rho$ growing from zero. Thus small $\rho$ interpolating solutions can be found before large $\rho$ interpolating solutions (associated with NTK). Thus for best classification a small initialization is required here as in the linear case, though the reason is quite different.

This note is a preliminary communication with more details, proofs and experiments to follow in a future update. We first provide the equations describing the dynamics of GD under a few different assumptions. A second part will describe our conjecture. Details and experiments are collected in the Appendix.

2 The dynamics of GD in $\rho$ and $V_k$

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 $L = \frac{1}{N} \sum_n \ell_n^2 = \frac{1}{N} \sum_n (g(x_n) - y_n)^2$, with $y = \pm 1$. In this note, we consider gradient descent instead of stochastic gradient descent.

Gradient descent on $L = \frac{1}{N} \left( \sum_n g_n^2 - 2 \sum_n y_n g_n + N \right)$ (using $g_n = g(x_n)$) gives

$$\dot{W}_k = -\frac{2}{N} \sum_n (g_n - y_n) \frac{\partial g_n}{\partial W_k}$$

that is

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

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

---

\[^3\text{Minimizing the norm minimizes stability}\]
2.2 Notation and assumptions

- We define \( g(x) = \rho f(x) \) with \( \rho \) defined as the products of the Frobenius norms of the weight matrices of the \( L \) layers in the network and \( f \) as the corresponding network with normalized weight matrices (because the RELU is homogeneous \(^6\));
- in the following we use the notation \( f_n \) meaning \( f(x_n) \), that is the normalized network;
- we assume \( \|x\| = 1 \) implying \( \|f(x)\| \leq 1 \) at convergence;

2.3 Dynamics under weight normalization

Gradient descent on \( L = \mathcal{L} + \lambda \sum_k V_k^2 = \sum_n (\rho f_n - y_n)^2 + \lambda \sum_k \|V_k\|^2 \) with \( \|V_k\|^2 = 1 \) is equivalent to “Weight Normalization”\(^6\) for deep networks. This dynamics can also be written as \( \dot{\rho}_k = V_k^T \dot{W}_k \) and \( \dot{W}_k = \rho S \dot{W}_k \) with \( S = I - V_k V_k^T \). This shows that if \( W_k = \rho_k V_k \) then \( \dot{V}_k = \frac{1}{\rho_k} \dot{W}_k \) as mentioned in \(^7\).

For networks with BN we assume (see \(^7\) ) \( \rho_k = 1, \forall k < L \) and \( \rho_L = \rho \) where \( L \) is the number of layers\(^4\) Then the dynamics of \( \rho = \rho_L \) if \(^5\)

\[
\dot{\rho} = -2[\sum_n \rho(f_n)^2 - \sum_n f_n y_n] = -2 \sum_n \ell_n f_n. \tag{3}
\]

where \( \ell_n = \rho f_n - y_n = \rho y_n f_n - 1 \) and\(^6\)

\[
\dot{V}_k = 2 \sum_n [(\rho f_n - y_n)\rho(-\frac{\partial f_n}{\partial V_k}) + 2V_k \rho f_n (\rho f_n - y_n)] = 2\rho \sum_n [(\rho f_n - y_n)(V_k f_n - \frac{\partial f_n}{\partial V_k})] \tag{5}
\]

**Equilibrium values**

The equilibrium value at \( \dot{\rho}_k = 0 \) is for all cases

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

Observe that \( \dot{\rho} = 0 \) if \( y_n f_n = 1 \) and \( \rho_k = 1; \rho_{eq} \) in Equation \(^6\) is a critical point for the dynamics of \( \rho \) under GD. In the case of SGD the asymptotic value of \( \rho \) for fixed \( \sum f_i y_i \) may fluctuate randomly around the \( \frac{\sum_n y_n f_n}{\sum_n f_n} \). Also observe that \( \dot{\rho}_k > 0 \) as long as \( \rho \) is not much larger than 1 and \( \sum_n f_n y_n > 0 \). Furthermore, the lowest value of \( \rho_k \) at equilibrium (\( \dot{\rho}_k = 0 \)) is \( \rho_k = 1 \)

\(^4\)Without BN \( \frac{\partial g_n(W)}{\partial W_k} = \frac{\rho}{\rho_k} \frac{\partial f_n(V)}{\partial W_k} \); with BN this becomes \( \frac{\partial g_n(W)}{\partial W_k} = \rho \frac{\partial f_n(V)}{\partial W_k} \), \( \forall k < L \) and \( \frac{\partial g_n(W)}{\partial W_L} = \frac{\partial f_n(V)}{\partial W_L} \).

\(^5\)In the appendix we describe a similar dynamics without the weight normalization algorithm.

\[
\dot{V}_k = -\frac{\partial L}{\partial V_k} = -2 \sum_n (\rho f_n - y_n) \rho \frac{\partial f_n}{\partial V_k} - 2\lambda V_k \tag{4}
\]

Because of the constraint imposed via Lagrange multipliers \( \|V_k\|^2 = 1 \), \( V_k^T \dot{V}_k = 0 \), which gives \( \lambda = -\sum_n (\rho^2 f_n^2 - \rho y_n f_n) \).
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 \( V_k = 0 \): they are minimizers with zero square loss. Notice that, in general, classification with maximum margin is not minimum norm interpolation of the labels.

In the case of \( V_k \) the learning rate is different for WN vs. no-WN but the equilibrium values are again the same. At equilibrium – that is when \( V_k = 0 \) – we have

\[
\sum_n (\rho f_n - y_n) \frac{\partial f_n}{\partial V_k} = \sum_n (\rho f_n - y_n) (V_k^{eq} f_n). \tag{7}
\]

Notice that \( \dot{\rho} = 0 \) implies \( \sum \ell_n f_n = 0 \). Then \( \dot{V}_k = 0 \) implies \( \sum \ell_n \frac{\partial f_n}{V_k} = 0 \); this latter condition implies the first one (multiply both sides by \( V^T \)).

Equation (7) suggests the following form for \( V_k^{eq} \) close to the minimum (when \( \dot{V}_k = \epsilon_7 \))

\[ V_k^{eq} = \sum \alpha_n \frac{\partial f_n}{\partial V_k}. \tag{8} \]

Notice that in general not all of the \( N \) terms in Equation (8) are different from zero or independent of each other (an upper bound is set by the rank of Jacobian \( \frac{\partial f_n}{V_k} \)). As an example consider the degenerate linear case when \( f \) is a linear function.

### 2.4 Weight decay

Adding a term \( \lambda |W_k|^2 \) to the Lagrangian corresponding to weight decay will change the dynamics of \( \rho \) but not the dynamics of \( V_k \). The Equation for \( \rho \) becomes

\[
\dot{\rho}_k = -2 \rho_k^{L-1} \sum_n \rho_k^L (f_n)^2 - \sum_n f_n y_n - 2\lambda \rho \tag{9}
\]

and

\[
\dot{\rho} = -2 \sum_n \rho (f_n)^2 - \sum_n f_n y_n - 2\lambda \rho \tag{10}
\]

with an equilibrium given by

\[
\rho_{eq} = \frac{\sum_n y_n f_n}{\lambda + \sum_n f_n^2}. \tag{11}
\]

Notice that label noise – suggested by Lee – may play a role somewhat similar to \( \lambda \) by making \( \sum_n f_n y_n \) smaller and thus also decreasing the equilibrium \( \rho \).

---

\(^7\)Since assuming that \( \sum \ell_n \neq zero \) we obtain \( V_k^{eq} = \sum \frac{\partial f_n}{\partial \ell_n} \).

\(^8\)Equation (8) has the same general form as provided by the Fritz-John lemma for the solution of a constrained optimization problem.
3 Linear thoughts

Consider two different linear problems where the interpolating minimizers of the square loss are respectively $F^a(x) = \rho_a f^a(x) = \rho_a v^a_T x$ and $F^b(x) = \rho_b f^b(x) = \rho_b v^b_T x$ with $\|v^a\| = \|v^b\| = 1$. Let us assume that $\rho_a < \rho_b$.

We then know (from Lemma 15 in Akshay+Lorenzo+Poggio) that the upper bound on the CVloo stability for $f_b$ is $\rho_b$ and for $f_a$ is $\rho_a$. Thus the bound on stability and expected error of $f_a$ vs. $f_b$ is $\frac{\rho_a}{\rho_b}$ times smaller.

4 Generalization in deep networks: a proposal

The logical steps in our theoretical proposal are

1. for each interpolating solution – corresponding to a global minimum $W^*$ – to which GD converges, there is a solution $W(X)$ which depends on the pseudoinverse of the Jacobian of $f$ wrt $W$ at $W^*$;

2. global minima with smaller $\rho$ have a larger number of values of $y_i f_i$ close (relatively) to 1, therefore with “largish” average margin;

3. margin bounds[8] suggests better expected error for large margin vs low margin interpolating solutions.

4.1 Degenerate minima, rank and minimum norm

Consider a K-layers network $f(W;x_i)$ parametrized by its weights $W$.

We first consider a simple setup with simplifying assumptions. We discuss later how to weaken the requirements.

Consider the zero-loss, interpolating equations

$$F(X,Y,W) = 0$$

where $Z = X,Y$ is the training set, $W$ is the set of weights across layers and $F(X,Y,W)$ is a set of $N$ equations, one for each of the $N$ data points (columns of $X$ and $Y$), that is $F_j(X,Y,W) = f(W;x_j) - y_j, \ j = 1, \cdots, N$. We assume $X \in \mathbb{R}^{N_x}$, $Y \in \mathbb{R}^{N_y}$ and $W \in \mathbb{R}^D$.

Under assumptions of differentiability of $F$, Equation 12 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. We state here an implicit function theorem for selections (reformulated from [9]) that does not require the Jacobian to be invertible

**Theorem 1** Consider the function $F : X \times W \rightarrow Y$, that is $F : \mathbb{R}^{N_x} \times \mathbb{R}^D \rightarrow \mathbb{R}^{N_y}$ with $N_x = N_y = N < D$ along with the associated solution mapping

$$S : X \rightarrow W \in \mathbb{R}^D |F(W,Z) = 0.$$
Let $F(W,Z) = 0$ so that $W^* \in S(Z^*)$. Assume that $F$ is strictly differentiable at $(X^*, W^*)$ and suppose that the partial Jacobian $\nabla_W F(X^*, W^*)$ is of rank $M \leq N$. Then the mapping $S$ has a “local selection” $s$ around $X^*$ for $W^*$ which is strictly differentiable with Jacobian $\nabla s(X) = A^T(AA^T)^{-1}\nabla_X f$, where $A = \nabla_W F(W^*, X^*)$.

Notice that Brower’s invariance of domain theorem says that the inverse $A^{-1}$ since $N_y < D$ fails to have a localization which is single-valued. However, though multivalued, $A^{-1}$ may contain a function with the property of single-valued localization for the case $N_y = D$. Such functions are called “selections”. In our case, $\nabla_X \rho$ depends only on a subset of $W$. However, their presence induces convergence to a minimum norm solution to happen faster than or WN are not required for this to happen in the classification case of exponential loss functions; however their presence induces convergence to a minimum norm (pseudoinverse) solution because $\rho$ are called “selections”.\footnote{In the section on minimizing the norm we show that when we minimize the norm subject to margin constraints, $W$ depends only on a subset of $X$, the so called support vectors that are on the margin.}

The relevance of the theorem for this paper is the statement about the existence of a differentiable function $W(x_1, \ldots, x_n)$ around the minimum $W^*$ and about its rank $M$, which depends on the rank of the Jacobian $\frac{\partial f}{\partial W}$.

Without WN or BN, gradient descent may converge to the degenerate minimum $W^*$ but in general without converging to the minimum norm (pseudoinverse) solution because $W$ will contain non-zero components outside the range of $A^T$. They correspond to non-zero initial conditions for convergence determined by the Jacobian. However, normalization algorithms such as WN or BN ensure that GD will converge to the local, linear pseudoinverse (see Appendix). Notice that BN or WN are not required for this to happen in the classification case of exponential loss functions; however their presence induces convergence to a minimum norm solution to happen faster than the asymptotic effect of the exponential loss.

Notice that the rank of the product of two matrices is the minimum of the two ranks: this helps interpreting the implications of the statement in the theorem that $\nabla s(X) = A^T(AA^T)^{-1}\nabla_X f$. Notice also that the rank of $\nabla s(X)$ is an upper bound on the $CV_{10}$ stability of the network since it corresponds to the number of independent training data to which $W^*$ sensitive. Thus the larger is the number of $y_n f_n$ that are significantly larger than zero, the better is the stability\footnote{For hard margin linear support vector machines we conjecture that the minimal number of SVs must be smaller than the rank of the data matrix}. Finally notice that if GD converges then $V_k = \frac{\partial f}{\partial V_k} = \nabla_{V_k} f$ (see Appendix).

### 4.2 Smaller $\rho$, larger margin

The first observation is that if we assume that the loss $L = \frac{1}{N} (\sum f^2_n - 2 \sum y_n f + N)$ is a continuous function of the $V_k$, then it will have at least one minimum at any fixed $\rho$, because its domain $V_k$ is compact. This means that for each $\rho$ there is at least a critical point for $V_k$.

Consider then such a critical point for which $V_k = 0$. Clearly critical points of $\rho$ do not exist if $\rho$ is too small.\footnote{Since usually the maximum output of a network depending on the number of layers is $<< 1$ the first critical point for increasing $\rho$ will be when $\rho$ is large enough to allow the following equation to have solutions}
\[ \sum_n y_n f_n = \rho \sum_n f_n^2. \] (14)

For increasing \( \rho \) we expect many solutions under GD\(^{12} \). The existence of many solutions is related to arguments showing the existence of NTK-based solutions: intuitively the last layer is enough in an extreme case – given enough overparametrization in the previous layer – to provide solutions for any set of random weights in the previous layers (for large \( \rho \) and small \( f_i \)). Furthermore the intermediate layer do not change much under GD in the iterations immediately after initialization. The emerging picture is a landscape in which there cannot be interpolating minima for \( \rho < 1 \) and in practice for significantly larger \( \rho \) than 1. With increasing \( \rho \) from \( \rho = 0 \) there will be a global degenerate (see \([10]\) ) minimum with the minimizer representing an interpolating solution with relatively large \( f_i \). Notice that the equilibrium value of \( \rho \) is a measure of “sparsity”: small \( \rho \) corresponds to \( f_i \) being close to either 1 or zero (\( \rho \) is in the order of \( \frac{1}{T} \)).

To give an intuition of what we expect qualitatively, given the previous analysis, let us assume that the initial conditions are \( \rho_t = 0 \approx 0 \) (with positive \( \rho_t = 0 \)), the network has more than two layers 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 does not go to zero), but very slowly for a longish time until it grows more quickly\(^{13} \).

Compare this with the case in which \( \rho \) is larger at initialization: then \( \rho \) grows very quickly to large values and then possibly decrease somewhat. As we already noticed, there are plenty of critical points for \( V_k \) at large \( \rho \). For larger and larger \( \rho \) there are an increasing number of critical points (stationary points for \( \rho \) and \( V_k \)). Notice that a small \( \rho \) may be seen as an implicit constraint corresponding to

\[ \min_{W_k} L = \sum (\rho f(x_i) - y_i)^2 \quad \text{s.t.} \quad \rho \leq C \] (15)

which is Ivanov regularization, which is itself “equivalent” to Tikhonov regularization with appropriate \( \lambda(\rho) \).

Our claim is that since \( \rho \) always increases, after separability is reached, up to an equilibrium value, small initialization allow the gradient flow to find minima with small \( \rho \) (and large margin), whereas large initialization makes this much more difficult because of the many minima encountered for decreasing \( \rho \) at large \( \rho \). Once a small \( \rho \) minimum is found, then the behavior with BN or WN is similar to iterative gradient descent in the case of linear networks (see for instance \([11]\) ) with convergence to the local minimum norm solution\(^{14} \).

\(^{12}\) It is interesting to recall \([10]\) that for SGD – unlike GD – the algorithm will stop only when \( \ell_n = 0 \quad \forall n \), which is the global minimum and corresponds to perfect interpolation. For the other critical points for which GD will stop, SGD will never stop but just fluctuate around the critical point.

\(^{13}\) Part of this behavior can be explained by the logistic equation in which the coefficients change with time, with \( y_n f_n \) decreasing slowly. As a consequence, the rate of increase of \( \rho \) decreases, though the asymptotic value of \( \rho = \frac{\sum y_n f_n}{\sum f_n^2} \) increases.

\(^{14}\) In the nonlinear case, the same behavior could be expected once GD is close to a minimum because around a degenerate minimum (local or global), 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 scenario above is as follows. Assume that we have separation and that $\rho$ is small. $\rho$ will eventually grow, until $\sum \rho f^2_n = \sum n y_n f_n$. Consider instead the case of a large $\rho$ initialization: $\rho$ will decrease until $\sum \rho f^2_n = \sum n y_n f_n$. There are many $V_k$ such that $f_n$ satisfying the condition are now easy to find. Of course, they are not as good in the sense that the $f_n$ are smaller and so the margins $y_n f_n$.

4.3 Stability and expected error

Assuming that small initialization provides a bias towards solution with “large” margin, the next step is to use simple bounds \cite{12} to claim better expected error (and better stability) for those solutions.

A typical generalization bound that holds with probability at least $(1 - \delta)$, $\forall f \in \mathbb{F}$ has the form \cite{12}:

\[
|L(f) - \hat{L}(f)| \leq c_1 \mathbb{R}_N(\mathbb{F}) + c_2 \sqrt{\frac{\ln(\frac{1}{\delta})}{2N}} \tag{16}
\]

where $L(f) = \mathbb{E}[\ell_{\text{gamma}}(f(x), y)]$ is the expected loss, is the empirical loss, $\mathbb{R}_N(\mathbb{F})$ is the empirical Rademacher average of the class of functions $\mathbb{F}$ measuring its complexity; $c_1, c_2$ are constants that reflect the Lipschitz constant of the loss function and the architecture of the network. The loss function here is the ramp loss $\ell_{\text{gamma}}(f(x), y)$ defined as

\[
\ell_{\text{gamma}}(y, y') = \begin{cases} 
1, & \text{if } yy' \leq 0, \\
1 - \frac{yy'}{\gamma}, & \text{if } 0 \leq yy' \leq \gamma, \\
0, & \text{if } yy' \geq \gamma. 
\end{cases}
\]

We define $\ell_{\text{gamma}=0}(y, y')$ as the standard 0−1 classification error and observe that $\ell_{\text{gamma}=0}(y, y') < \ell_{\text{gamma}>0}(y, y')$.

As in the linear case we now consider two solutions with zero empirical loss of the square loss regression problem obtained with the same RELU deep network and corresponding to two different minima with two different $\rho$s. Let us call them $\rho_a f^a(x)$ and $\rho_b f^b(x)$. Using the notation of this paper, the functions $f_a$ and $f_b$ correspond to networks both with normalized weight matrices at each layer.

Let us assume that $\rho_a < \rho_b$.

We now use the observation that, because of homogeneity of the networks, the empirical Rademacher complexity satisfies the property,

\[
\mathbb{R}_N(\mathbb{F}) = \rho \mathbb{R}_N(\tilde{\mathbb{F}}), \tag{17}
\]

where $\mathbb{F}$ is the space of functions of our unnormalized networks and $\tilde{\mathbb{F}}$ denotes the corresponding normalized networks.\footnote{Furthermore, the Rademacher complexity of the space of functions associated with normalized networks of the same architecture is the same (see \cite{8}).} This observation allows us to use the bound Equation 16 and the
fact that the empirical $L_\gamma$ for both functions is zero (if $\gamma < \rho_a$) to write \( L_0(f^a) = L_0(F^a) \leq c_1\rho_a R_N(F) + c_2 \sqrt{\frac{\ln(\frac{1}{\delta})}{2N}} \) and \( L_0(f^b) = L_0(F^b) \leq c_1\rho_b R_N(F) + c_2 \sqrt{\frac{\ln(\frac{1}{\delta})}{2N}} \). The bounds have the form

\[
L_0(f^a) \leq A\rho_a + \epsilon \quad (18)
\]

and

\[
L_0(f^b) \leq A\rho_b + \epsilon \quad (19)
\]

Thus the bound for the expected error $L_0(f^a)$ is better than the bound for $L_0(f^b)$.

Remarks It would be interesting to avoid the use of the margin bounds and connect directly the lower margin (and large norm) solutions to lower stability. A possible way to proceed is to use the argument in section 3 applying it to the Jacobian $\nabla s(X)$.

Another approach would be to connect the regression setup of this paper with a constrained optimization in which results such as the Fritz John lemma could be applied.

Notice that stability (and expected error) is upper bounded by the rank of the Jacobian at $W^*$ which is at most $N$. Equation 8 shows that $V_k$ depends on the $\frac{\partial f^a}{\partial V_k}$ with $M$ non-zero coefficients, where $M$ is related to the rank of the Jacobian of $f(X; V)$ wrt $V_k$. In particular, if $M = N_y$ that is the number of data points, the $CV_{loo}$ stability of such a solution will be poor. On the other hand, if $M << N_y$, then we expect good stability and low expected error.

Notice also is that $\rho$ is by itself a measure of complexity of the hypothesis space; interestingly small $\rho$ roughly correlates with large “average” margin here that is with $\sum y_n f_n \sum f_n^2$.

5 Summary

- Convergence to a local minimum norm solution is expected when normalization techniques such as Batch Normalization (BN) or Weight Normalization (WN) are used, in a way which is similar to the behavior of linear degenerate networks under GD.

- The bias towards large margin interpolating solutions (of minimum norm) depends on zero-initial conditions. Thus zero initial conditions induce GD to explore first the small norm, large margin solutions and only later the large norm, small margin ones. The expected error of the minimum norm solution is bounded by the rank of the local Jacobian $\frac{\partial f^a}{\partial V_k}$ (at the minimum $W^*$), associated with the pseudoinverse in the solution. More relevant is that margin bounds relate large margins to better bounds on the expected error.

There are open questions, one of which is as follows. Under the assumption of early separability during the first iterations of GD, small initialization ensures that $\rho$ grows for small values thus exploring first large margin minima. Why does GD have difficulties in converging in the absence of BN, especially for very deep networks? At the moment, the best answers are that

\[16\] Every choice of a leave-one-out example will induce a change in $V_k$

\[17\] The former is the case of CiFAR with random labels; the latter is the case of CIFAR with natural labels.
- BN is needed to ensure convergence to the minimum local norm
- good tuning of the learning rate is important and BN ensures autotuning [7].

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


A Un-normalized GD

Here we assume (for simplicity of derivations) that at initialization all the layers have the same norm, that is \( \rho_k \) is the same for all \( k \) at initialization. Because of this assumption we can use the following

**Lemma 2** \( \frac{\partial \rho^2_k}{\partial t} \) is independent of \( k \).

To claim that all \( \rho_k \) are the same at all times. Thus \( \rho = \rho^L_k \), where \( L \) is the number of layers.

**Proof** Consider \( \frac{\partial ||W_k||^2}{\partial t} \). The calculation follows the case for the exponential loss:

\[
\frac{\partial ||W_k||^2}{\partial t} = 2W_k \frac{\partial W_k}{\partial t} = \frac{4}{N} \sum_n g_n^2 + \frac{4}{N} \sum_n g_n \tag{20}
\]

because of the structural lemma. Thus the time evolution of \( \rho^2 = ||W_K||^2 \) is independent of \( k \).

Then we obtain the dynamical system

\[
\dot{\rho}_k = -\frac{\partial L}{\partial \rho_k} = -2 \sum_n (\rho_k^L f_n - y_n) f_n \rho_k^{L-1} = -2 \rho_k^{L-1} \left[ \sum_n \rho_k^L (f_n)^2 - \sum_n f_n y_n \right] \tag{21}
\]

which can be rewritten in terms of \( \rho = \rho^L_k \) using \( \dot{\rho} = \sum_k \frac{\partial \rho}{\partial \rho_k} \dot{\rho}_k \) as

\[
\dot{\rho} = 2 \rho^L \frac{2L-2}{L} \left[ \sum_n f_n y_n - \sum_n \rho (f_n)^2 \right] \tag{22}
\]

which is an equation of the type known as “differential logistic equation” used for instance to model sigmoidal population growth. It has an interesting dynamics as shown in the simplified simulations in the appendix. This is not the case for the BN dynamics of Equation 3.

Figure 1: Binary classification on two classes from CIFAR-10, trained with MSE loss. The model is a very simple network with 4 layers of fully-connected Layers. ReLU nonlinearity is used. The weight matrices of all layers are initialized with zero-mean normal distribution, scaled by a constant such that the Frobenius norm of each matrix is either 5, 15 or 30. We run SGD with batch size 128, constant learning rate 0.1 and momentum 0.9 for 1000 epochs. No Batch normalization. No weight decay. No data augmentation. Every input to the network is scaled such that it has Frobenius norm 1.
Figure 2: Dynamics of $\rho$ from experiments in Figure 1. First row: small initialization (5). Second row: large initialization (15). Third row: extra large initialization (30). A dashed rectangle denotes the previous subplot’s domain and range in the new subplot. More details to be added.
Figure 3: Margin of all training samples

Figure 4: Results with a convolution network and batch normalization: Binary classification on two classes from CIFAR-10, trained with MSE loss. The model is a very simple network with 5 layers of convolutions. ReLU nonlinearity is used. Batch normalization is used without parameters (affine=False in PyTorch). The weight matrices of all layers are initialized with zero-mean normal distribution, scaled by a constant such that the Frobenius norm of each matrix is either 0.1 or 5. We run SGD with batch size 128, constant learning rate 0.01 and momentum 0.9 for 1000 epochs. No weight decay. No data augmentation. Every input to the network is scaled such that it has Frobenius norm 1.
Figure 5: Results with a convolution network and batch normalization: Dynamics of $\rho$ from experiments in Figure 4. Top row: small initialization (0.1). Bottom row: large initialization (5). The plot starts with $\rho(0) = 0$ despite an initialization of $\rho_k = 0$ because the scaling factor of BN starts from 0. A dashed rectangle denotes the previous subplot’s domain and range in the new subplot.
Figure 6: Results with a convolution network and batch normalization: Margin of all training samples

The dynamics of Equation 22 is that the smaller ρ₀=0 is, the longer it takes to ρ to grow (this phenomenon increase with larger number of layers L). Thus ρ is constrained by the nonlinear dynamics to be very small for a transient phase T of GD iterations (as we mentioned, T is longer with more layers and longer with smaller initialization) and then to grow slowly while ƒn grows towards 1 (implying that ∑ ƒn² approaches ∑ yn ƒn).

B Linear networks with Weight Normalization

Consider the separable case of a linear network (ƒ(x) = ρvTx) the dynamics is with ℓn = e−ρynvTxn for the exponential loss and ℓn = (ρvTxn − yn) for the square loss

\[ \dot{\rho} = \frac{1}{\rho} \sum_{n=1}^{N} \ell_n y_n v^T x_n \]  

(23)

and

\[ \dot{v} \propto \sum_{n=1}^{N} \ell_n (x_n - vv^T x_n). \]  

(24)

Since elln ≠ 0 in general, let us assume that ∑n=1N ℓnyn ≠ 0. Thus

\[ \dot{v} \propto \sum_{j=1}^{N} \alpha_j (I - vv^T) x_j. \]  

(25)
Figure 7: Results with a convolution network and batch normalization: Histogram of $|f_n|$ over time. Top figure: initial $\rho_k = 0.1$. Bottom figure: initial $\rho_k = 5$. 
If gradient descent converges to \( \dot{v} = 0 \), the solution \( v \) must satisfy \( vv^T x = x \), where \( x = \sum_{j=1}^{N} \alpha_j x_j \). Assume \( \|x\| = 1 \). The operator \( T \) in \( v(t+1) = T v(t) \) associated with equation [25] is not expanding because \( v \) has unit norm. Thus [13] there is a fixed point \( v = x \) which is independent of initial conditions.

B.1 Remarks

- The condition number of \( Ax = y \) is defined considering how much a relative change in \( y \) changes \( x \) in relative terms. It is thus independent of the norm of \( y \). The absolute condition number is different and depends on \( y \). It may be relevant in the presence of noise with power independent of norms. For instance for linear regression \( V^T X = \frac{1}{\rho} Y \).

- Equation 6 is a critical point for the dynamics of \( \rho \) under GD but NOT under SGD. In the case of SGD the asymptotic value of \( \rho \) for fixed \( \sum f_i y_i \) can 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 \( V_k \).

- 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.

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

- Suppose we control \( \rho_k \) independently of \( V_k \) and of equation 22, can we find \( \rho(t) \) schedules leading to solutions with better generalization?

- Notice that in the case of BN \( \frac{\partial g_n(W)}{\partial W_k} = \rho \frac{\partial f_n(V)}{\partial V_k} \) for \( k \neq L \) and \( \frac{\partial g_n(W)}{\partial W_L} = \frac{\partial f_n(V)}{\partial V_L} \) otherwise. This suggests that for all layers apart the final one, the Jacobians are \( \frac{\partial f_n(V)}{\partial V_k} = \frac{1}{\rho} \frac{\partial g_n(W)}{\partial W_k} \).

- It seems that square loss reaches solutions with good test error in multiclass CIFAR10 faster than cross-entropy. Continuing GD, however, yields overfitting for the square loss (and worse test error) but not for cross-entropy. This is very interesting because it validates the asymptotic complexity control we described in [14]. This suggests that in the experiments of Belkin et al. early stopping may play a role to obtain results with the square loss case that are as good or better than cross-entropy.

- small \( \rho \) seems to imply sparsity in the example set, that is \( W_k = \sum_{n \in I} \alpha_n \frac{\partial f_n}{\partial W_k} \) where \( I \) is the subset of \( n \) with minimum margin = 1; the latter implies good test error via stability (see [3]) if \( I \) is small;
• Predicting from training data the test error: one way is to what about comparing two solutions by seeing how much you have to multiply the one with the smaller margin (a factor $\alpha > 1$) to get the same margin as the one with the bigger one. This should correspond to the ratio of the norms of the unnormalized solutions. To compute the margin one can take the worse point in the training set or perhaps the average margin of the same number of worse points for the two solutions.

We consider the dynamical system induced by GD on a deep net with

C Maximum margin and minimum norm

Lemma 3 \[6\] The maximizer of the margin when $||V_k|| = 1$ is the minimum norm solution when the margin is $\geq 1$.

D Nonlinear networks with Weight Normalization

Consider

$$\dot{V}_k = \alpha \sum_{n=1}^{N} e^{-\rho y_n f_n} y_n \left( \frac{\partial f_n}{\partial V_k} - V_k V_k^T \frac{\partial f_n}{\partial V_k} \right) = (I - V_k V_k^T) \sum_{n=1}^{N} e^{-\rho y_n f_n} y_n \frac{\partial f_n}{\partial V_k}. \quad (26)$$

The equation becomes, defining $\alpha_j = y_j e^{-\rho y_j f_j}$ and $\frac{\partial f}{\partial V_k} = \sum_j \alpha_j \frac{\partial f_j}{\partial V_k}$, $(I - V_k V_k^T) \frac{\partial f}{\partial V_k} = \dot{V}_k$. If there is convergence that is $\dot{V}_k = 0$ then

$$V_k = \frac{\partial f}{\partial V_k}. \quad (27)$$

does that mean that the class means before the last layer become the final classifier weights? Notice that the solution is given in an implicit way assuming convergence since $f$ depends on $V_k$.

If there is convergence at a finite time to a ball around the minimum $V_k$ then we can assume that $-\epsilon \leq \dot{V}_k \leq \epsilon$ where the bound is meant componentwise and $\epsilon > 0$.

E Normalized representation and NTK

The conditions that yield a NTK linear approximation correspond to minimizing $L = \sum_n (f_n - \frac{y_n}{\rho})^2$ with large $\rho$ so that the targets are close to 0. Under this parametrization $f$ can be well approximated by a linear function in the $V_k$. Under other conditions, there is a convergence of $\sum_n (f_n - \frac{y_n}{\rho})^2$ under the dynamics \[15\].
F In general classification with maximum margin is not minimum norm interpolation of the labels (Lorenzo)

Minimum norm regression of binary labels is

$$\min_{W_k} \frac{1}{2} \|W_k\|^2, \quad \forall k \text{ subj. to } y_i f(W_K, \cdots, W_1; x_i) = 1, \quad i = 1, \ldots, N. \quad (28)$$

Minimum norm binary classification is

$$\min_{W_k} \frac{1}{2} \|W_k\|^2, \quad \forall k \text{ subj. to } y_i f(W_K, \cdots, W_1; x_i) \geq 1, \quad i = 1, \ldots, N. \quad (29)$$

Clearly classification involves minimizing over a larger class of functions than regression. The result will be in general different.

Observation 1 Minimum norm binary classification under the square loss with margin 1 (implying $f_V(x_i) \geq 1 \forall i$) is not (in general) interpolation of all the data.

Notice that hard margin SVM is a case in point: the SVs interpolate their data point, but other non-support vectors have margin greater than one.

G Equivalence of Tikhonov, Morozov and Ivanov regularization

Consider the following minimization problems

1. $$\min_W L(W) \quad s.t. R(W) \leq \delta \quad (30)$$

2. $$\min_W R(W) \quad s.t. L(W) \leq \epsilon \quad (31)$$

3. $$\min_W L(W) + \lambda R(W) \quad (32)$$

Under mild conditions the three minimization problems are equivalent. For instance, Ivanov’s is equivalent to Tikhonov in the sense that for a given $\lambda$ and data set, there exist a value of $\delta$ such that the two solutions are the same [16]. TO BE DONE FORMALLY

Notice that we may replace $\min_W L(W)$ in the minimization problems above with $\min_{V, \rho} L(V, \rho)$.

To show that Ivanov implies Tikhonov for the ridge regression problem, we need to demonstrate strong duality and that the dual optimum is attained. Both of these things are implied by Slater’s constraint qualifications. Show that the Ivanov form of ridge regression is a convex optimization problem with a strictly feasible point.
**H Stability and compression**

Let us estimate algorithmic stability at a minimum of the loss (we write $f_V$ as $f$):

$$\forall S \in \mathbb{Z}^n, \quad \forall i = 1, \cdots, n, ||A_S(x) - A_{S'}(x)||_\infty$$

(33)

We estimate in expectation

$$(E_S|A_S(x) - A_{S'}(x)|^2 \leq E_S||A_S(x) - A_{S'}(x)||^2$$

(34)

Assume that the minimum solution only depends on a single training point $x_*$. We estimate

$|A_S(x) - A_{S'}(x)| = |f^*(x) - f^*(x)| = 0$ when $x_i \neq x_*$ and $|A_S(x) - A_{S'}(x)| \leq 2$ when $x_i = x_*$. In expectation this gives $\beta_C = \frac{2}{n}$. More in general the following holds

**Theorem 4 (informal)** Suppose there are $h$ support vectors on which the minimizer depend . Then the bound on stability is $\frac{2h}{N}$.

Notice that the previous informal theorem can be viewed from the point of view of information compression which is known to be closely connected to generalization. Notice also that one would of course expect a significant difference in $h$ between CIFAR with random labels and CIFAR with natural labels.