On the Learning Dynamics of Deep Neural Networks

Remi Tachet des Combes¹, Mohammad Pezeshki¹,², Samira Shabanian¹, Aaron Courville², Yoshua Bengio²
¹Microsoft Research Montreal
²Université de Montréal, MILA

Abstract

While a lot of progress has been made in recent years, the dynamics of learning in deep nonlinear neural networks remain to this day largely misunderstood. In this work, we study the case of binary classification and prove various properties of learning in such networks under strong assumptions such as linear separability of the data. Extending existing results from the linear case, we confirm empirical observations by proving that the classification error also follows a sigmoidal shape in nonlinear architectures. We show that given proper initialization, learning expounds parallel independent modes and that certain regions of parameter space might lead to failed training. We also demonstrate that input norm and features' frequency in the dataset lead to distinct convergence speeds which might shed some light on the generalization capabilities of deep neural networks. We provide a comparison between the dynamics of learning with cross-entropy and hinge losses, which could prove useful to understand recent progress in the training of generative adversarial networks. Finally, we identify a phenomenon that we baptize gradient starvation where the most frequent features in a dataset prevent the learning of other less frequent but equally informative features.

1 Introduction

Due to extremely complex interactions between millions of parameters, nonlinear activation functions and optimization techniques, the dynamics of learning observed in deep neural networks remain much of a mystery to this day. What principles govern the evolution of the neural network weights? Why does the training error evolve as it does? How do data and optimization techniques like stochastic gradient descent interact? Where does the implicit regularization of deep neural networks trained with stochastic gradient descent come from? Shedding some light on those questions would make training neural networks more understandable, and potentially pave the way to better techniques.

It is commonly accepted that learning is composed of alternating phases: plateaus where the error remains fairly constant and periods of fast improvement where a lot of progress is made over the course of few epochs (Saxe et al., 2013a). Theoretic explanations of that phenomenon exist in the case of regression on linear neural networks (Saxe, 2015) but extensions to the nonlinear case (Heskes & Kappen, 1993; Raghu et al., 2017; Arora et al., 2018) fail to provide analytical solutions.

It has been observed in countless experiments that deep networks present strong generalization abilities. Those abilities are however difficult to ground in solid theoretical foundations. The fact that deep network have millions of parameters – a number sometimes orders of magnitude larger than the dataset size – contradicts the expectations set by classic statistical learning theory on the necessity of regularizers (Vapnik, 1998; Poggio et al., 2004). This observation drove Zhang et al. (2016) to suggest the existence of an implicit regularization happening during the training of deep neural networks. Advani & Saxel (2017) show that the dynamics of gradient descent can protect against overfitting in large networks. Kleinberg et al. (2018) also offer some explanations of the phenomenon but understanding its roots remains an open problem.
In this work, we study the learning dynamics of a deep nonlinear neural network trained on a standard classification task using two different losses: the cross-entropy and the hinge loss. We mainly focus on binary classification, some of the results and properties can however be extended to the multi-class case. The questions we address in Sections 3, 4 and 5 respectively can be summarized as follows:

- How does the confidence of a classifier evolve throughout learning?
- How does the loss used during training impact its dynamics?
- Which properties of the features present in a dataset impact learning, and how?

**Independent mode learning** We show that, similarly to the case of linear networks and under certain initial conditions, learning happens independently between different classes, i.e., classes induce a partition of the network activations, corresponding to orthogonal modes of the data.

**Learning dynamics** We prove that in accordance to experimental findings, the hidden activations and the classification error of the network show a sigmoidal shape with slow learning at the beginning followed by fast saturation of the curve. We also characterize a region in the initialization space where learning is frozen or eventually dies out.

**Hinge loss** We study how using the hinge loss impacts learning and quantitatively compare it to the classic cross-entropy loss. We show that the hinge loss allows one to solve a classification task much faster, by providing strong gradients no matter how close to convergence the neural network is.

**Gradient starvation** Finally, we identify a phenomenon that we call gradient starvation where the most frequent features present in the dataset starve the learning of other very informative but less frequent features. Gradient starvation occurs naturally when training a neural network with gradient descent and might be part of the explanation as to why neural networks generalize so well. They intrinsically implement a variant of Occam’s razor (Ariew, 1976): the simplest explanation is the one they converge to first.

### 2 Setup and notations

We are interested in a simple binary classification task, solved by training a deep neural network with gradient descent. This simple setup encompasses for instance the training of generative adversarial networks discriminators. Some of our results extend to multi-class classification, but, for the sake of conciseness, that case is only treated in the Appendix. We let $D = \{(x_i, t_i)\}_{1 \leq i \leq n} \subset \mathbb{R}^d \times \{1, 2\}$ denote our dataset of vectors and labels. The classifier we consider is a simple neural network with one hidden layer of $h$ neurons and a ReLU non-linearity (see Fig. 1).

The output of the network is passed through a function denoted $o$ which is either the sigmoid $\sigma$ in the binary cross-entropy case or the identity in the hinge loss case. The full function can be written as

$$P_t(x) := o(u_t(x)) := o(Z_t^T W_t x_),$$

where $W_t$ is an $h \times d$ matrix and $Z_t$ a vector of length $h$. In the $C$-class case, $Z_t$ is instead a $C \times h$ matrix and $o$ the softmax function. We will see below an extension to deeper networks. $W_t$ and $Z_t$ are the parameters of the neural network. The subscript $+$ (resp. $t$) denotes the positive part of a real number (resp. the state of the element at time step $t$ of training). The superscript $T$ stands for the transpose operation. In the case of cross-entropy, $P_t(x)$ represents the probability that $x$ belongs to class 1, for the hinge loss, $P_t(x)$ is trained to reach $\{1, -1\}$ for classes 1 and 2. For a given class $k$, we let $D_k$ denote the set of vectors belonging to it. We make the assumption

(H1) For any $x, x' \in D_k$, $x^T x' > 0$. For any $x \in D_1$ and $x' \in D_2$, $x^T x' \leq 0$.

It implies linear separability of the data, an assumption often necessary in theoretical studies (Soudry et al., 2017; Liao & Couillet, 2018; Nacson et al., 2018; Xu et al., 2018) and the positioning of the origin between the two sets. It is a very strong assumption which admittedly bypasses a large part of the deep learning dynamics. Nevertheless, it allows the discovery of interesting properties and helps develop intuitions about behaviors observed in more general settings.
3 Learning dynamics for binary cross-entropy

In this section, we focus on the case of the binary cross-entropy loss

$$L_{BCE}(W_t, Z_t; x) = -\mathbb{1}_{x \in D_1} \log(\sigma(Z^T_t(W_t x)_+)) - \mathbb{1}_{x \in D_2} \log(1 - \sigma(Z^T_t(W_t x)_+)),$$

and train our network using stochastic gradient descent to minimize $L_{BCE}$.

3.1 Independent modes of learning

Our first lemma states that over the course of training and under suitable initialization, the active neurons of the hidden layer remain the same for each datapoint, and the coordinates of $Z_t$ remain of the same sign. To prove it, we let $w^i_t$ denote the $i$-th row of $W_t$ and make the additional assumptions: there exists a partition $\{I_1, I_2\}$ of $\{1, \ldots, h\}$ such that with $k \in \{1, 2\}$

(H2) For any $i \in I_k, x \in D_k$ and $x' \notin D_k, w^i_k x > 0$ and $w^i_k x' \leq 0$.

(H3) The $i$-th coordinate of $Z_0$ is positive if $i \in I_1$, negative otherwise.

Assumption [H2] states that at the beginning of training, data points from different classes do not activate the same neurons. It is an analogue to the orthogonal initialization used in [Saxe et al. 2013b]. In Appendix A.4 we show that relaxing it hints towards an extended period of slow learning in the early stages of training. [H3] is introduced for Lemma 3.1 and Theorem 3.2 but will be relaxed later.

Lemma 3.1. For any $k \in \{1, 2\}$, $x \in D_k$ and $t \geq 0$, the only non-negative elements of $W_t x$ are the ones with an index $i \in I_k$. The signs of the coordinates of $Z_t$ remain the same throughout training.

This lemma proves that updates to the parameters of our network are fully decoupled from one class to the other. An update for a data point in $D_1$ will only influence the corresponding active rows and elements of $W_t$ and $Z_t$. This “independent mode learning” is an equivalent of the results by Saxe et al. (2013b) in a non-linear network trained on the cross entropy loss. The proof of the lemma and its extension to $N - 1$ hidden layers and multi-class classification can be found in Appendix A.

3.2 Learning dynamics

We are now interested in the actual dynamics of learning, and move from discrete updates to continuous ones by considering an infinitesimal learning rate $\alpha$ [Heskes & Kappen, 1993]. Lemma 3.1 can easily be extended to this setting. For simplicity we assume $h = 2$, but similar results hold for arbitrary $h$ (see Appendix A.4). For the moment, we maintain the assumptions [H1][H3]

Theorem 3.2. Assuming that each class $k$ contains the same vector repeated $|D_k|$ times, then the output of the classifier on that class converges at a speed proportional to both $|D_k|/|D|$ and to the norm of the vector it contains. The shape of the learning dynamics is shown on Fig. 2.

Proof. To simplify the notations, we arbitrarily assume that $I_1 = \{1\}$ and we write $w_t$ and $z_t$ the row and element modified by an update made using $x \in D_1$ (the case of $D_2$ can be treated symmetrically). By the independence proven above, we know that $w_t$ and $z_t$ are only affected by updates from $D_1$. This greatly simplifies our evolution equations to: $w'_t = \delta_f(x)z_t x^T$ and $z'_t = \delta_f(x) w_t x$ where $\delta_f(x)$ is the gradient of the loss with respect to the pre-sigmoid output of the network $u_t(x)$: $\delta_f(x) = 1_{\{k=1\}} - \sigma(u_t(x))$ and the prime indicates a time derivative. We let $y_t = w_t x$, which gives

$$y'_t = \frac{x^T x}{1 + \epsilon y_t z_t}, \quad z'_t = \frac{y_t}{1 + \epsilon y_t z_t}. \quad (1)$$

Writing $\|x\|^2 = x^T x$, we see that the quantity $y^2_t - \|x\|^2 z^2_t$ is an invariant of the problem, so its solutions live on hyperbolas of equation $y^2 - \|x\|^2 z^2 = \pm c$ with $c := |y^2_0 - \|x\|^2 z^2_0|$.

We only treat the case of a degenerate hyperbola $c = 0$ i.e. $y^2_0 = \|x\|^2 z^2_0$, and refer the interested reader to Appendix A.3 for the full derivation. In the case $c = 0$, we have $\forall t, \ y_t = \|x\|z_t$ (those quantities are both positive as $x \in D_1$ and [H2][H3]). $u(t) := z_t w_t x = z_t y_t$, thus follows the equation $u'(t) = 2\|x\| u(t) \sigma(-u(t))$. One can see a clear equivalence between our evolution equation and the one obtained in Saxe et al. (2013b). Its analytical solution is (see Appendix A.3)

$$u(t) = (\log + Ei)^{-1} (2\|x\| t + \log(u_0) + Ei(u_0)), \quad (2)$$

3
Formally, this allows us to quantify the performance of the network at any time $D$. Each couple lives on a hyperbola. The slope of the linear curves is equal $\pm \|x\|$ (set to 0.7 in this diagram). The green region represents the initializations of $(y, z)$ where the classification task will be solved by the network. In the red region, learning does not start (the neuron is not active initially) or collapses as the neuron dies off when $y$ reaches 0. The $c_i$ points represent the three cases from Section 3.3 Right. Plots of $y_t$ and $P_t(x \in D_1) = \sigma(z_t y_t)$ for different values of $c$ and $\|x\|$.

where $E_i$ is the exponential integral (Wiki., 2018) and $<+1>$ denotes the inverse function.

We let $\bar{u}$ denote that function for $\|x\| = 1$. For $x \in D_2$, the degeneracy assumption becomes $y_0 = -\|x\|$. It can be shown similarly that $v(t) := z_t y_t$ verifies the equation $v'(t) = 2\|x\|v(t)\sigma(v(t))$ with a negative initial condition (H2-3) In other words, $u$ and $v$ follow symmetric trajectories on the positive/negative real line. Below, $\bar{u}$ (resp. $v$) denote those two trajectories for $\|x\| = 1$ and initial conditions $u_0 > 0$ (resp. $v_0 < 0$).

Let us now write $p = |D_1|/|D|$ the fraction of points in the dataset belonging to $D_1$. Because we sample randomly from the dataset, this amounts to sampling $p$ (resp. $1-p$) points from $D_1$ (resp. $D_2$) for each time unit during training, i.e. to rescaling the time axis by $p$ for $D_1$ and $1-p$ for $D_2$. Formally, this allows us to quantify the performance of the network at any time $t$

$$
\begin{cases}
    P_t(x \in D_1) = \bar{u}(\|x\|pt) & \text{if } x \in D_1 \\
    P_t(x \in D_2) = \bar{u}(-\|x\|((1-p)t)) & \text{if } x \in D_2
\end{cases}
$$

(3)

In particular, the convergence of $u(t)$ to $+\infty$ can be bounded using our results: convergence happens at a rate slower than $\log(t)$ (Appendix A.3, a fact proved on its own by Soudry et al. [2017]).

**Interpretation** Fig. 2 Right. shows the learning dynamics for different values of $\|x\|$ and $c$. One common characteristic between all the curves is their sigmoidal shape. Learning is slow at first, then accelerates before saturating. This is aligned with empirical results from the literature. We also see on e.g. the blue and yellow curves that a larger $\|x\|$ (or similarly a larger $p$) converges much faster. The effect of $c$ on the dynamics can mostly been seen at the beginning of training (for instance on the green and yellow curves). It fades as convergence happens, corresponding to points of the hyperbolas getting closer to the asymptote $y = \|x\|z$, see Fig. 2 Left. and below for more details. We can characterize the convergence speeds more quantitatively with the

**Corollary 3.3.** Let $\delta$ be the required accuracy on the classification task (i.e. $P_A(x \in D_1) \geq 1 - \delta$ for $x \in D_1$). Under certain assumptions, the times $t_1^*$ and $t_2^*$ required to reach that accuracy for each classifier verify $t_2^* \approx t_1^* \frac{\|x_k\|}{\|x_k\| - \|x_k\|_{L_1}} p$ where $\|x_k\|$ is the norm of vector $\|x_k\|$ from class $k$.

The proof can be found in Appendix A.5. More frequent classes and larger inputs will be classified at a given level of confidence faster. The class frequency observation is fairly straightforward as updates on a more frequent class occur at a higher rate. As far as input sizes are considered, this can be seen as an analogous to the results from Saxe et al. (2013b) stating that input-output correlations drive the speed of learning. Because a sigmoid is applied on the network output, its (pre-sigmoid) targets are sent to $+\infty$. A larger input is more correlated with its target and converges faster.
On the assumptions The assumption that each class only contains one vector allows us to obtain the first closed-form solutions of the learning dynamics for the binary cross-entropy. It can be relaxed to classes containing orthogonal datapoints (see Appendix A.8) which still remains restrictive. A possible interpretation is the following: if one were to consider a deep neural network that has learnt two discriminative features for the two classes, applying classic SGD on those features would result in a learning rate proportional to the prominence of those two features in the original dataset, and to learning curves of that exact shape. It is worth noting that such shapes are regularly observed by machine learning practitioners (Saxe et al., 2013b), our results help better understand them.

3.3 Phase diagram

In this section, we build the phase diagram of Fig. 2 Left. The notations follow Theorem 3.2 in particular $y_t = w_1 x$. So far, we have considered points in the top-right quadrant. In that region, the couple $(z_t, y_t)$ lives on a hyperbola of equation $y^2 - ||x||^2 z^2 = \pm c$ where $c \geq 0$. The sign in the equation is defined by the position of $(z_0, y_0)$ relative to the function $y = ||x|| z$ (positive if above, negative otherwise). If $(z_0, y_0)$ is originally on that line, it will remain there throughout training.

We now explore the rest of the parameter space by relaxing some of our assumptions. We still consider a point $x \in D_1$, the diagram for $D_2$ can be obtained by mirroring the $z$ axis.

**On Assumption [H2]** Let us first consider the simple case of $w_0 x = y_0 < 0$. The neuron is initially inactive because of the ReLU. No updates will ever be made to $w_1$ during training. This corresponds to the bottom half of the phase diagram, the parameters are frozen (also see Advani & Saxe, 2017).

**On Assumption [H2]** We now assume that $z_0 \leq 0$. In that case, a simple extension of Lemma 3.1 shows that learning still happens independently on each row of $w_1$. The outcome from Theorem 3.2 is still valid: the couple $(z_t, y_t)$ lives on a hyperbola. It is however not guaranteed anymore that $y_t$ shall remain positive throughout training. There are three possible situations (numbered 1 to 3), each represented by the corresponding point on the diagram.

1) If $y_0 = - ||x|| z_0$, then the points $(z_t, y_t)$ are stuck in the top-left quadrant and converge to zero. The equation verified by the logit $u(t)$ is $u'(t) = -2 ||x|| u(t) \sigma(-u(t))$ (see Appendix A.7).

2) If $y_0 > - ||x|| z_0$, the points $(z_t, y_t)$ move on the hyperbola towards the top-right quadrant, at which point $z_t$ becomes positive and $y_t$ starts increasing again.

3) If $y_0 < - ||x|| z_0$, the points $(z_t, y_t)$ move on the hyperbola towards the bottom-left quadrant, at which point $y_t$ becomes negative. When that happens, the neuron dies out, learning stops.

Only in the second case will the classifier end up solving the task: random initialization functions in certain parts of the $(y_t, z_t)$ space. Those findings are summarized in the phase diagram Fig. 2 Left. The red region represents the initialization where the network will not be able to solve the task.

3.4 Deeper neural networks

In a similar fashion to the above, we can extend our analysis to a network with $N - 1$ hidden layers of the same shape. Under some stringent assumptions on the initialization of the network, the classes are learnt independently from one another. We can also show that the ordinary differential equation verified by the logit $u(t)$ of our network is

$$u'(t) = \frac{N ||x||^{2/N} u^{2-2/N}(t)}{1 + e^{u(t)}}.$$  \hspace{1cm} (4)

The full analysis can be found in Appendix B. Solutions of Eq. 4 are plotted in Fig. 3.

Here too, a sigmoidal shape appears during the learning process. The effect of $||x||$ reduces as $N$ grows due to the power $2/N$, however, larger values still converge faster (e.g. the blue and yellow
For Theorem 4.1, in Theorem 3.2 (some of which can be relaxed, see Appendix C), we have the following result:

The hinge loss is non differentiable, but one can simply consider that learning stops as soon as the features in the case where $c = 0$, leading to the result. When $c \neq 0$, a classic hyperbolic change of variables allows to find the solution. Its full derivation can be found in Appendix [C].

The learning curves are plotted in Fig. [4] Right. We notice a *hard* sigmoidal shape corresponding to learning stopping when $u_t$ reaches 1. The comparison to binary cross-entropy (all other parameters kept equal) suggests a much faster convergence. With $\delta$ the required confidence for our classifier, the time $t^*$ required to reach $\delta$ can easily be computed. We plot it for both losses in Fig. [4] Right which confirms visually that the hinge loss converges much faster. We also notice the expected divergence of $t^*$ for the binary cross entropy as $\delta$ reaches 1 (training never converges in that case). We refer the interested reader to Appendix [C] for a more general treatment of the Hinge loss, fully relaxing the assumption on the number of points in the classes.

5 Gradient Starvation

In this section, we attempt to quantify the impact of feature frequency inside a given class. We keep our simplified framework and consider that the input $x$ to our network is composed of two underlying features $x_1 \in \mathbb{R}^{d_1}$ and $x_2 \in \mathbb{R}^{d_2}$ with $d = d_1 + d_2$. We let $(x_1, x_2) \in \mathbb{R}^d$ denote the concatenation of the vectors $x_1$ and $x_2$. We assume that all the points in class $D_1$ contain the feature $x_1$ but only a
fraction $\lambda$ of them contains the feature $x_2$. This is equivalent to making continuous gradient updates using the vector $(x_1, x_2)$ with a rate $\lambda$ and the vector $(x_1, 0)$ with a rate $1 - \lambda$. We also assume that those features are fully informative for $D_1$—i.e., are absent from class $D_2$. A network trained using gradient descent on the dataset we just described has the following property:

Even though the feature represented by $x_2$ is fully informative of the class, the network will not classify a sample containing only $x_2$ with high confidence.

It is the result of a phenomenon we coin gradient starvation where the most frequent features starve the gradient for the least frequent ones, resulting in a slower learning of those:

**Theorem 5.1.** Let $\delta$ be our confidence requirement on class $D_1$ i.e. training stops as soon as $\forall x \in D_1, P_t(x \in D_1) \geq 1 - \delta$, and let $t^*$ denote that instant. Then, under some mild assumptions,

$$P_t^*((0, x_2) \in D_1) \leq \frac{1}{1 + e^{-\lambda \log(\frac{1 - \delta}{\lambda})}}.
$$

**Proof.** From Lemma 3.1, assumptions (H2-3) are sufficient to guarantee independent mode learning as well as positiveness of $z_t$. We decompose $w_t = (\alpha t x_1, \beta_t x_2) + (x_1^T, x_2^T)$ where $x_1, x_1^T = x_2^T = 0$, and assume that $\alpha_0 \geq \beta_0 > 0$ (in App. D we relax some of those assumptions and prove an equivalent result). The evolution equation for $w_t$ is $w_t' = \frac{x_t}{1 + e^{z_t}}$ with $x = (x_1, x_2)$ (resp. $(x_1, 0)$) at an $\lambda$ (resp. $1 - \lambda$) rate. Projecting on $x_1$ and $x_2$ gives

$$\alpha_t' = \lambda \frac{z_t}{1 + e^{z_t}}, \quad \beta_t' = \lambda \frac{z_t}{1 + e^{z_t}}.
$$

From $z_t > 0$, we see that $\beta_t$ is an increasing function of time, which guarantees $\beta_t > 0$, and $\alpha_t' \geq \beta_t'(1 - \lambda) = (1 + \frac{1 - \lambda}{\lambda})\beta_t' = \frac{\beta_t'}{\lambda}$.

This proves that $\forall t \geq 0$, $\alpha_t \geq \beta_t/\lambda$. We now consider $t^*$ such that $z_{t^*} \alpha_{t^*} = \log(\frac{1 - \delta}{\lambda})$. It is the smallest $t$ such that $P_t((x_1, x_2) \in D_1) \geq P_t((x_1, 0) \in D_1) = 1 - \delta$, its existence is guaranteed by $z_t$ and $\alpha_t$ being increasing (we also assume that $t = 0$ does not verify those inequalities). We get

$$P_{t^*}((0, x_2) \in D_1) = \frac{1}{1 + e^{-z_{t^*} \beta_{t^*}}} \leq \frac{1}{1 + e^{-\lambda z_{t^*} \alpha_{t^*}}} = \frac{1}{1 + e^{-\lambda \log(\frac{1 - \delta}{\lambda})}}.
$$

As can be seen in Eq. 7, the presence of $\alpha_1$—which detects feature $x_1$—in the denominator of $\beta_1'$ greatly reduces its value, thus preventing the network from learning $x_2$ properly. □

| Table 1: Gradient starvation |
|-----------------------------|
| $\delta$ | $\lambda$ |
| 0.5     | 0.99% | 91% | 71% | 61% | 51% |
| 0.2     | 99%   | 96% | 86% | 72% | 52% |
| 0.1     | 99.99%| 98% | 86% | 72% | 52% |
| 0.01    | 99.99%| 98% | 86% | 72% | 52% |

| Table 2: Accuracy on the cats and dogs dataset (Real means the untouched test set) |
|-----------------------------|
| Training | Testing | Testing (Real) |
| 100%    | 100%    | 43.2%          |

In Fig. 5 Left., we plot for different values of $\lambda$ the confidence of the network when classifying $x_2$ as a function of its confidence on $x_1$ (see App. D for more details). The gap between the two is very significant: with e.g. $\lambda = 0.1$ and $\delta = 10^{-4}$ (Table 1), $P_{t^*}((0, x_2) \in D_1) \leq 72\%$! Even though $x_2$ is exclusively present in $D_1$, and is thus extremely informative, the network is unable to classify it.
Experiment To validate those findings empirically, we design an artificial experiment based on the cats and dogs dataset (Kaggle, 2018). We create a very strong, perfectly discriminative feature by making the dog pictures brighter, and the cat pictures darker. We then train a standard deep neural network to classify the modified images and measure its performance on the untouched testing set.

Results The results can be seen in Table 2. The network perfectly learns to classify both the train and test modified set, but utterly fails on the real test data. This proves that the handcrafted light feature was learnt by the network, and is used exclusively to classify images. All the features allowing to recognize a cat from a dog are still present in the data, but the low level features (e.g. the presence of whiskers, how edges combine to form the shape of the animals and so on) are far less frequent than the light intensity, and thus were not learnt. The most frequent feature starved all the others.

6 Related work

The learning dynamics of neural networks have been explored for decades. Baldi & Hornik (1989) studied the energy landscape of linear networks and the fixed point structure of gradient descent learning in that context. Heskes & Kappen (1993) developed a theory encompassing stochastic gradient descent and parameter dynamics and wrote down their evolution equations in on-line learning. However, those equations are heavily nonlinear and do not have closed-form solutions in the general case. Saxe et al. (2013b) study the case of deep linear networks trained with regression. They prove the existence of nonlinear learning phenomena similar to those seen in simulations of nonlinear networks and provide exact solutions to the dynamics of learning in the linear case. Some of our results are an extension of theirs to nonlinear networks. Choromanska et al. (2014); Raghu et al. (2017); Saxe (2013); Yosinski et al. (2014) also focus on neural network dynamics, while Nacson et al. (2018); Xu et al. (2018); Soudry et al. (2017) study the convergence rate of learning on separable data. Arora et al. (2018) prove that overparameterization can lead to faster optimization.

Recent work in the domain of generative adversarial networks (Goodfellow et al., 2014) has shown the resurgence of the hinge loss (Rosasco et al., 2004). In particular, part of the success encountered by Miyato et al. (2018) is due to their use of that specific loss function. Their main contribution however is a spectral normalization technique that produces state-of-the-art results on image generation. Their paper is part of a larger trend focusing on the spectra of neural network weight matrices and their evolution during learning (Vorontsov et al., 2017; Odena et al., 2018; Pennington et al., 2017). It, nevertheless, remains a poorly understood subject.

Zhang et al. (2016) performed some experiments proving that deep neural networks expound a so-called implicit regularization. Even though they have the ability to entirely memorize the dataset, they still converge to solutions that generalize well. A variety of explanations for that phenomenon have been advanced: correlation between flatness of minima and generalization (Hochreiter & Schmidhuber, 1997), natural convergence of stochastic gradient descent towards such minima (Kleinberg et al., 2018), built-in hierarchical representations (LeCun et al., 2015), gradient descent naturally protecting against overfitting (Advani & Saxe, 2017), and structure of deep networks biasing learning towards simpler functions (Neyshabur et al., 2014; Perez et al., 2018). Our results from Section 5 suggest that gradient descent indeed has a beneficial effect, but can also hurt in some situations.

7 Discussion

In order to obtain closed form solutions for the learning dynamics, we made the extremely simplifying assumption that each class only contains one point. We leave overcoming that limitation to future work. In the spirit of the proof in Section 5 where we considered two datapoints, we might be able to obtain upper and lower bounds on the learning dynamics.

Our comparison between the cross-entropy and the hinge losses reveals fundamental differences. It is noteworthy that the hinge loss is an important ingredient of the recently introduced spectral normalization (Miyato et al., 2018). The fast convergence of networks trained with the hinge loss might in part explain its beneficial impact. A deeper analysis of the connections between the two would lead to a better understanding of the performance of the algorithm.

In this paper, we introduce the concept of gradient starvation and suggest that it might be a plausible explanation for the generalization abilities of deep neural networks. By focusing most of the learning
on the frequent features of the dataset, it makes the network ignore the idiosyncrasies of individual datapoints. That rather desirable property has a downside however: very informative but rare features will not be learnt during training. This strongly limits the ability of the network to transfer to different data distributions where e.g. the rare feature exists on its own.

REFERENCES

Madhu S. Advani and Andrew M. Saxe. High-dimensional dynamics of generalization error in neural networks. CoRR, abs/1710.03667, 2017.

Roger Ariew. Ockham’s Razor: A Historical and Philosophical Analysis of Ockham’s Principle of Parsimony. PhD thesis, University of Illinois at Urbana-Champaign, 1976.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. CoRR, abs/1802.06509, 2018. URL http://arxiv.org/abs/1802.06509

P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. Neural Networks, 2:53–58, 1989.

Anna Choromanska, Mikael Henaff, Michaël Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surface of multilayer networks. CoRR, abs/1412.0233, 2014. URL http://arxiv.org/abs/1412.0233

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672–2680, 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

Tom M. Heskes and Bert Kappen. On-line learning processes in artificial neural networks, 1993.

S. Hochreiter and J. Schmidhuber. Flat minima. Neural Computation, 9(1):1–42, 1997.

Kaggle. Dogs vs. cats, 2018. URL https://www.kaggle.com/c/dogs-vs-cats

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL http://arxiv.org/abs/1412.6980 cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? Corr, abs/1802.06175, 2018. URL http://dblp.uni-trier.de/db/journals/corr/corr1802.html#abs-1802-06175

Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. Nature, 521(7553):436–444, 2015. URL http://dblp.uni-trier.de/db/journals/nature/nature521.html#LeCunBH15

Zhenyu Liao and Romain Couillet. The dynamics of learning: A random matrix approach. In Jennifer G. Dy and Andreas Krause (eds.), ICML, volume 80 of JMLR Workshop and Conference Proceedings, pp. 3078–3087. JMLR.org, 2018. URL http://dblp.uni-trier.de/db/conf/icml/icml2018.html#LiaoC18a

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks, 2018. arXiv:1802.05957v1.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. CoRR, abs/1803.01905, 2018. URL http://dblp.uni-trier.de/db/journals/corr/corr1803.html#abs-1803-01905

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. CoRR, abs/1412.6614, 2014. URL http://dblp.uni-trier.de/db/journals/corr/corr1412.html#NeyshaburTS14
Augustus Odena, Jacob Buckman, Catherine Olsson, Tom B. Brown, Christopher Olah, Colin Raffel, and Ian J. Goodfellow. Is generator conditioning causally related to gan performance? CoRR, abs/1802.08768, 2018. URL http://dblp.uni-trier.de/db/journals/corr/corr1802.html#abs-1802-08768

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In NIPS-W, 2017.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, Roman Garnett (eds.), NIPS, pp. 4788–4798, 2017. URL http://dblp.uni-trier.de/db/conf/nips/nips2017.html#PenningtonSG17

Guillermo Valle Perez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. CoRR, abs/1805.08522, 2018. URL http://dblp.uni-trier.de/db/journals/corr/corr1805.html#abs-1805-08522

Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi. General conditions for predictivity in learning theory. Nature, 428(6981):419–422, 2004.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep understanding and improvement. arXiv preprint arXiv:1706.05806, 2017.

Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same?. Neural Computation, 16(5):1063–107, 2004. URL http://dblp.uni-trier.de/db/journals/neco/neco16.html#RosascoVCPV04

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Learning hierarchical categories in deep neural networks. In Markus Knuff, Michael Pauen, Natalie Sebant, and Ipke Wachsmuth (eds.),CogSci. cognitivesciencesociety.org, 2013a. ISBN 978-0-9768318-9-1. URL http://dblp.uni-trier.de/db/conf/cogsci/cogsci2013.html#SaxeMG13

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. cite arxiv:1312.6120, 2013b. URL http://arxiv.org/abs/1312.6120

Andrew Michael Saxe. Deep Linear Neural Networks: A Theory of Learning in the Brain and Mind. PhD thesis, Stanford University, 2015.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. CoRR, abs/1710.10345, 2017. URL http://dblp.uni-trier.de/db/journals/corr/corr1710.html#abs-1710-10345

Vladimir Naumovich Vapnik. Statistical Learning Theory. Wiley-Interscience, September 1998.

Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Chris Pal. On orthogonality and learning recurrent networks with long term dependencies. CoRR, abs/1702.00071, 2017. URL http://arxiv.org/abs/1702.00071

Wiki. Exponential integral, 2018. URL https://en.wikipedia.org/wiki/Exponential_integral

Tengyu Xu, Yi Zhou, Kaiyi Ji, and Yingbin Liang. Convergence of sgd in learning relu models with separable data. CoRR, abs/1806.04339, 2018. URL http://dblp.uni-trier.de/db/journals/corr/corr1806.html#abs-1806-04339

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Advances in neural information processing systems, pp. 3320–3328, 2014.
Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016. URL [http://dblp.uni-trier.de/db/journals/corr/corr1611.html#ZhangBHRV16](http://dblp.uni-trier.de/db/journals/corr/corr1611.html#ZhangBHRV16).
A.1 Proof of Lemma 3.1

In this section, we prove the following lemma:

**Lemma 3.1** For any $k \in \{1, 2\}, x \in D_k$ and $t \geq 0$, the only non-negative elements of $W_t x$ are the ones with an index $i \in I_k$. The signs of the coordinates of $Z_t$ remain the same throughout training.

**Proof.** We prove this with a simple induction. The claim is true at $t = 0$ from assumptions [(H1-3)]. Let us assume that at time set $t$, the different parts of the lemma are true. The SGD updates to the weights stemming from a single observation $x \in D_k$ (the extension to a mini-batch is straightforward) with a learning rate $\alpha$ are:

$$
\Delta Z_t(x) = \alpha \delta_f(x) (W_t x)_+ \\
\Delta W_t(x) = \alpha \delta_f(x) (Z_t^T \circ e_k) x^T,
$$

where $\circ$ denotes the element-wise product of two vectors and $e_k \in \mathbb{R}^h$ is the binary vector with ones on the indices from $I_k$. $\delta_f(x)$ is the gradient of the loss with respect to the pre-sigmoid output of the network $F_t(x)$: $\delta_f(x) = 1_{\{x = 1\}} - \sigma(F_t(x))$. The updates to $Z_t$ are positive on its indices belonging to $I_k$, and negative otherwise, proving the second claim of the lemma. Moving to $W_t$, only its rows and hidden neurons with indices $i \in I_k$ are modified. For an element $x' \in D$, $W_{t+1} x' = W_t x' + \alpha \delta_f(x) (\sum_{i \in I_k} z_i) x^T$ where $z_i$ is the $i$-th element of $Z_t$. By induction, $\sum_{i \in I_k} z_i$ has the same sign as $\delta_f(x)$ (positive on $D_1$, negative on $D_2$). From [(H1)] $x^T$ is positive (resp. negative) if $x'$ is in $D_k$ (resp. otherwise). The update keeps the $k$-th neuron active (resp. inactive). \(\square\)

A.2 Extension of Lemma 3.1 to multi-class classification

Let us consider a simple $C$-classification task. $D = \{(x_i, y_i)\}_{1 \leq i \leq n} \subset \mathbb{R}^d \times C$ is our dataset of vectors/labels where $C$ denotes both the set of possible labels and its cardinal depending on the context. The classifier we are training is a simple neural network with one hidden layer of $C$ neurons, a ReLU non-linearity and a softmax $\sigma$. The full function is written as

$$
P_t(x) := \sigma(F_t(x)) := \sigma(Z_t(W_t x)_+),$$

where $W_t$ (resp. $Z_t$) is a $C \times d$ (resp. $C \times C$) weight matrix and $\sigma$ the softmax function. The subscript $+$ denotes the positive part of a real number. $P_t(x)$ is a $C$-vector representing the probability that $x$ belongs to each class. The subscript $t$ denotes the state of the element at time step $t$ of training.

For a given class $k$, we let $D_k$ be the set of vectors belonging to class $k$. We make the following assumptions, with $k \neq k' \in C$ two arbitrary classes:

(H4) For any $x, x' \in D_k$, $x^T x' > 0$. For any $x \in D_k$ and $x' \in D_{k'}$, $x^T x' \leq 0$.

(H5) For any $x \in D_k$ and $x' \in D \setminus D_k$, $w^k_0 x > 0$ and $w^k_0 x' \leq 0$, where $w^k_0$ is the $k$-th row of $W_0$.

(H6) $Z_0$ is initialized randomly to positive numbers on the diagonal, and non-positive elsewhere.

Those assumptions are straightforward extensions of the ones from the main text. We train the classifier using stochastic gradient descent on the cross-entropy loss of our problem. $F_{t+1}$ is the state of the neural network after one SGD update to $F_t$. Our first lemma states that over the course of training, the active neurons of the hidden layer remain the same for each element of the dataset, and the elements of $Z_t$ remain of the same sign.

**Lemma A.1.** For any $k \in C$, $x \in D_k$ and $t \geq 0$, the only non-negative element of $W_t x$ is its $k$-th element and all diagonal (resp. non-diagonal) elements of $Z_t$ are positive (resp. negative).

**Proof.** We prove this with a simple induction. The claim is true at $t = 0$ from assumptions [(H4-6)]. Let us assume that at time set $t$, the different parts of the lemma are true. The SGD updates to the weights stemming from a single observation $x \in D_k$ (the extension to a mini-batch is straightforward) and a learning rate $\alpha$ are:

$$
\delta Z_t = \alpha \nabla_y L (W_t x)_{+}^{T} \\
\delta W_t = \alpha (Z_t^{T} \nabla_y L \circ e_{k}) x^{T},
$$

12
where \( \otimes \) denotes the element-wise product of two vectors and \( e_{k^*} \) the \( k^* \) basis vector of \( \mathbb{R}^C \). \( \nabla_y L \) is the gradient of \( \log F_i(x) \), \( u \) is a rescaling of \( \mathbb{R}^+ \rightarrow [u_0, +\infty] \) (since its derivative is strictly positive). We have

\[
t = \frac{1}{2||x||} \int_{u_0}^{u_f} \frac{1}{y} \frac{dy}{d ||x||} = \frac{1}{2||x||} \left[ \log(u_f) + Ei(u_f) - Ei(u_0) \right],
\]

where \( Ei(x) = -\int_{-\infty}^{x} \log(-u) du \) is the exponential integral. In the end, with the superscript \(<-1>\) denoting the inverse function

\[
u_f = u(t) = (\log + Ei)^{<-1>}(2||x||t + \log(u_0) + Ei(u_0)).
\]

We let \( \bar{u} \) denote the function \( u(t) \) above for \( ||x|| = 1 \), the solution for \( ||x|| \neq 1 \) can easily be inferred by a rescaling of \( t \) in \( \bar{u} \). For \( x \in D_2 \), the system of ODEs is

\[
y'_t = -\frac{||x||^2 z_t}{1 + e^{ynz_t}}, \quad z'_t = -\frac{yt}{1 + e^{ynz_t}}.
\]
the degeneracy assumption becomes \( y_0 = -\|x\|z_0 \) (we know from \([\text{H2}]\) that \( y_0 > 0 \) and from \([\text{H3}]\) that \( z_0 < 0 \)). It can be shown similarly that \( v(t) := z_ty_t \) verifies the equation \( v'(t) = 2\|x\|\sigma(v(t)) \) with a negative initial condition. In other words, \( u \) and \( v \) follow symmetric trajectories on the positive/negative real line. Below, \( \bar{u} \) and \( \bar{v} \) denote those two trajectories for \( \|x\| = 1 \) and initial conditions \( u_0 > 0 \) and \( v_0 < 0 \).

Let us now write \( p = |D_1|/|D| \) the fraction of points in the dataset belonging to \( D_1 \). Because we sample randomly from the dataset, this amounts to sampling \( p \) (resp. \( 1 - p \)) points from \( D_1 \) (resp. \( D_2 \)) for each time unit during training, ie to rescaling the time axis by \( p \) for \( D_1 \) and \( 1 - p \) for \( D_2 \).

Formally, this allows us to quantify the performance of the network at any time \( t \)

\[
\left\{ \begin{array}{ll}
P_t(x \in D_1) = \sigma(\bar{u}(\|x\|pt)) & \text{if } x \in D_1, \\
P_t(x \in D_2) = \sigma(-\bar{v}(\|x\|/(1 - p)t)) & \text{if } x \in D_2, \\
\end{array} \right.
\]

which concludes the proof for \( c = 0 \).

From Eq. \( [14] \) we also see that

\[
t = \frac{1}{2\|x\|} \int_{u_0}^{u_f} 1 + e^y \, dy \geq \frac{1}{2\|x\|} \int_{u_0}^{u_f} e^{\xi} \, dy = \frac{1}{\|x\|} (e^{u_f} - e^{u_0}),
\]

where we used the inequality \( \forall y \geq 0, e^{\xi} > y \). It eventually gives us

\[
u(t) \leq 2 \log(|\|x\|t + e^{u_0}|),
\]

a result in line with convergence rates obtained in \([\text{Soudry et al.} 2017]\).

Let us now study the non-degenerate case. We apply the classic change of coordinates

\[
y_t = \sqrt{c} \cosh(\frac{\theta}{2}), \quad z_t = \frac{\sqrt{c}}{\|x\|} \sinh(\frac{\theta}{2}) \quad \text{if } y_0^2 > \|x\|^2 z_0^2,
\]

\[
y_t = \sqrt{c} \sinh(\frac{\theta}{2}), \quad z_t = \frac{\sqrt{c}}{\|x\|} \cosh(\frac{\theta}{2}) \quad \text{if } y_0^2 < \|x\|^2 z_0^2.
\]

Since \( y_t^2 + \|x\|^2 z_t^2 = c \cosh(\theta) \) and \( y_t z_t = c/2\|x\| \sinh(\theta) \), we see that

\[
(y_t z_t)' = \frac{y_t^2 + \|x\|^2 z_t^2}{1 + e^{y_t z_t}} = \frac{c}{2\|x\|} \cosh(\theta) \theta' = \frac{c \cosh(\theta)}{1 + e^{c \sinh(\theta)/2\|x\|}},
\]

where the first equality used the system of equations Eq. \( [11] \). This gives us the dynamics of \( \theta \) as

\[
\theta' = \frac{2\|x\|}{1 + e^{c \sinh(\theta)/2\|x\|}},
\]

with an initial condition \( \theta_0 = \cosh^{-1}(y_0^2 + \|x\|^2 z_0^2 / c) \). For any \( \theta_f \geq \theta_0 \), we see that \( t = \theta_c^{-1}\theta(\theta_f) \) verifies

\[
t = \int_{\theta_0}^{\theta_f} \frac{1 + e^{c \sinh(\theta)/2\|x\|}}{2\|x\|} \, d\theta.
\]

There is no closed-form solution for that integral (that we know of). It can however be computed numerically. On Fig. 2 Right of the main text, we plot the curves for \( y_t \) and \( \sigma(z_t y_t) \) for different values of \( c \) and \( \|x\| \). We obtain a sigmoidal shape similar to previously made empirical observations. We notice in particular that for larger values of \( \|x\| \) the function converges faster.

\( \square \)

\textbf{A.4 \ Extension to \( h \) hidden neurons}

We now extend the result from \textbf{Theorem 3.2} to the case with \( h \) hidden neurons. Let us still consider an update made on \( x \in D_1 \). We know from assumptions and by \textbf{Lemma 3.1} that the only active neurons in the network are indexed by \( I_1 \). The network weights follow the evolution equations:

\[ (w_t^i)' = \frac{x^T z_t^i}{1 + e^{\sum_{j \in I_1} z_t^j w_t^j} x}, \quad (z_t^i)' = \frac{w_t^i x}{e^{\sum_{j \in I_1} z_t^j w_t^j} x}. \]
We similarly define $y_t^i = w_t^i x$ which brings

$$
(y_t^i)' = \frac{x^T x z_t^i}{1 + e^\sum_{j \in x_1} z_t^j y_t^j}, \quad (z_t^i)' = \frac{y_t^i}{e^\sum_{j \in x_1} z_t^j y_t^j}.
$$

The couple $(y_t^i, z_t^i)$ follows the same hyperbolic invariance, defined by a constant $c_i$. In the case where $\forall i \in I_1, c_i = 0$, we see that $u_t^i := z_t^i y_t^i$ verifies

$$
(u_t^i)' = \frac{2 \|x\| u_t^i}{1 + e^\sum_{j \in x_1} u_t^j},
$$

and a simple summation on $i$ shows that $u_t := \sum_{i \in x_1} u_t^i$ follows Eq. [13]. The dynamics of the logit in this case are identical to the single active neuron case, the only difference is potentially its initial value.

A.5 Proof of Corollary [3.3]

**Corollary 3.3** Let $\delta$ be the required accuracy on the classification task (i.e. $P_t(x \in D_1) \geq 1 - \delta$ for $x \in D_1$). Under certain assumptions, the times $t_1^*$ and $t_2^*$ required to reach that accuracy for each classifier verify

$$
t^* \approx \|x_1\| \|x_2\| \frac{p}{1-p} \text{ where } \|x_k\| \text{ is the norm of vector } \|x_k\| \text{ from class } k.
$$

**Proof.** We consider the case $c = 0$. The classification error drops at a rate proportional to $\|x_1\| p$ for $D_1$ and to $\|x_2\| (1 - p)$ for $D_2$. More precisely, let us assume that $u_0 \leq |v_0|$ and look for a classification confidence of $1 - \delta$. We write $u_f = \sigma^\alpha - \sigma^\beta$ $(1 - \delta) = \log(1/\delta - 1), t_v = u < -(u_0)$ and

$$
t^* = \frac{1}{2} \left( \log(1/\delta - 1) \right) + E_i(\log(1/\delta - 1)) - E_i(u_0),
$$

t^* \text{ represents the time taken to reach confidence } \delta \text{ with an initialization } u_0, \text{ and } t_v \text{ the time to reach } v_0 \text{ starting in } u_0. \text{ We see that}

$$
P(x \in D_1) \geq 1 - \delta \iff t \geq t_1^* = \frac{t^*}{\|x_1\| p} \text{ if } x \in D_1,
$$

$$
P(x \in D_2) \geq 1 - \delta \iff t \geq t_2^* = \frac{t^* - t_v}{\|x_2\| (1 - p)} \text{ if } x \in D_2.
$$

The ratio between the convergence times reads

$$
t_2^* = \frac{\|x_1\| p}{\|x_2\| (1 - p)} \text{ (1 - t_2^*)}. \text{ One sees that if the weight initializations } u_0 \text{ and } v_0 \text{ are close (i.e. } t_v \text{ is small) and the confidence requirement large (i.e. } t^* \text{ is large), the ratio is approximately } \frac{\|x_1\| p}{\|x_2\| (1 - p)}.
$$

A.6 Relaxing assumption (H2)

Let us first recall that the assumption states:

(H2) For any $i \in I_k, x \in D_k$ and $x' \notin D_k, w_0^i x > 0$ and $w_0^i x' \leq 0$.

We now assume that $h = 2$ and study the evolution of the first row of matrix $W_t$, written $w_t$ in the following. We relax assumption (H2) by assuming that there is a point $x_2$ in $D_2$ such that $w_0 x > 0$. And we consider updates to $w_t$ coming from sampling equally $x_1$ from $D_1$ and $x_2$. The evolution equations can be written as

$$
w_t^i = \frac{x_1^T x_t^i z_t}{1 + e^{z_t^i w_t x_1}} - \frac{x_2^T x_t^i z_t}{1 + e^{-z_t^i w_t x_2}}, \quad z_t^i = \frac{w_t x_1}{1 + e^{z_t^i w_t x_1}} - \frac{w_t x_2}{1 + e^{-z_t^i w_t x_2}}.
$$

In order to make the analysis simpler, we assume that $x_1^T x_2 = 0$. We decompose $w_t = \alpha_t x_1 + \beta_t x_2 + x'$ and get

$$
\alpha_t' = \frac{z_t}{1 + e^{z_t^i \alpha_t}}, \quad \beta_t' = -\frac{z_t}{1 + e^{-z_t^i \beta_t}}, \quad z_t^i = \frac{\alpha_t}{1 + e^{z_t^i \alpha_t}} - \frac{\beta_t}{1 + e^{-z_t^i \beta_t}}.
$$

From our relaxation of assumption (H2), we know that $\beta_0 > 0$. We also have $\alpha_0 > 0$ and $z_0 > 0$. This means that $\alpha_t$ is increasing, and that $\beta_t$ is decreasing until it reaches 0, at which point learning
in that direction stops. A competition appears between $\alpha_t$ and $\beta_t$ to determine the sign of $z'_t$. More importantly, $\alpha_t z_t$ which measures the confidence the network has in $x_1$ belonging to class 1 evolves according to

$$(\alpha_t z_t)' = \frac{z_t^2 + \alpha_t^2}{1 + e^{\alpha_t}} - \frac{\alpha_t \beta_t}{1 + e^{-z_t \beta_t}}.$$  

The negative term slows down the learning of $D_1$ until $\beta_t$ reaches 0. When that happens, we obtain the regime described in the main text. This observation might be a clue as to why in more complex networks with random initialization, the slow phase of learning can last for a long time: the network needs to build its independent modes of learning.

A.7 Top-left quadrant initialization

When the initial conditions of the network verify $y_0 = -\|x\| z_0$, then at all time $t$, $y_t = -\|x\| z_t$. Plugging that equality in Eq. [11] results in

$$u'(t) = \frac{-2\|x\| u(t)}{1 + e^{u(t)}}$$

where again $u(t) = z_t y_t$. This means that the logit is negative and increasing. Let $u_0 < u_1 < 0$, we see that the time at which $u$ reaches $u_f$ verifies

$$t = -\frac{1}{2\|x\|^2} \int_{u_0}^{u_1} \frac{1 + e^y}{y} dy = \frac{1}{2\|x\|^2} \int_{u_0}^{-u_f} \frac{1 + e^{-y}}{y} dy \geq \log(-u_0) - \log(-u_f).$$  

$t$ diverges to $+\infty$ as $u_f$ tends to 0 from below. The logit converges to 0 without ever reaching it.

A.8 Relaxing the single datapoint assumption

One of the major assumptions made in the main text is the fact that each class contains a single element. In this section, we slightly relax it to the case where the points $\{x_c\}_{1 \leq c \leq m} \subset \mathbb{R}^d$ in a class are all orthogonal to one another \footnote{This implies in particular $m \leq d$} (while still verifying Assumption [H1]). In that case, each presentation of a training vector will only affect $w_i$ in the direction of that specific vector. Let $y'_i$ denote $w_i x_i$, the unnormalized component of $w_i$ along $x_i$. We consider a batch update on the weights of the neural network. In that case:

$$(y'_i)' = \frac{\|x_i\|^2 y'_i}{1 + e^{\|x_i\| y'_i}}, \quad (z'_i)' = \sum_{i=1}^{m} \frac{y'_i}{1 + e^{\|x_i\| y'_i}}.$$  

Assuming that the vectors all have the same norm (denoted $\|x\|$ below) and that the $y'_i$ are all equal, then that equality remains true at all time (they follow the same update equation). We let $y_t$ denote that value:

$$(y'_t)' = \frac{\|x\|^2 z'_t}{1 + e^{\|x\| z'_t}}, \quad (z'_t)' = \frac{m y_t}{1 + e^{\|x\| y_t}}.$$  

A new invariant appears in those equations: $c := |m y_0^2 - \|x\|^2 z_0^2|$. In the case $c = 0$ (the other case can be treated as above), we obtain the following evolution equation for the logit of any point in the class:

$$u'(t) = \frac{2\sqrt{m} \|x\| u(t)}{1 + e^{u(t)}}.$$

We end up with a similar equation than before except for the $\sqrt{m}$ factor, which boosts the convergence speed. However, one should not forget that we are now training on a full batch (i.e. on $m$ points) during each unit of time. Performing the same number of updates for the single point class would generate a $m$ factor in the convergence speed of $u$ (one $\sqrt{m}$ factor for each $y$ and $z$ functions). The slower convergence for the more general case can be explained by the fact that each point is making an update on $w_i$ in its own direction. That direction being orthogonal to all others points makes it useless for their classification.
Appendix B.

In this section we study the case of a deeper network with $N - 1$ hidden layers. Similarly to above, we can prove the existence of independent modes of learning. To that end, neurons need to be activated in a disjoint manner from one class to the other. Due to the growing complexity of the interactions between the parameters of the network, this requires very strong assumptions on the initialization of the network and on the shape of the network. Assuming that the network is written as

$$P_t(x) := \sigma(Z_t^T (Z_t^{N-2} \cdots (Z_t^1(W_t x)_+ + \ldots )_+) ,$$

with $W_t$ an $h \times d$ matrix and for all $1 \leq i \leq N - 2$, $Z_t^i$ an $h \times h$ matrix. We maintain assumption (H2) from the main text, and extend (H3) to all $Z_t^i$ by assuming that they are diagonal, and that the $j$-th element of their diagonal is positive if $j \in I_1$, negative otherwise. To simplify notations, we go back to assuming $h = 2$ and take an update on $x \in D_1$. Only the first elements of every matrix are modified, we write them $z_t^i$ and keep the notations $z_t$, $w_t$ and $y_t$. We can then write the evolution equations:

$$z_t^i = \frac{z_t^{i-1} y_t}{1 + e^{u(t)}}, \quad (z_t^i)' = \frac{z_t^{i-1} y_t}{1 + e^{u(t)}}, \quad y_t' = \frac{\|x\|^2 z_t^{N-2} \cdots z_t^1}{1 + e^{u(t)}},$$

with $u(t) = z_t \Pi z_t^i y_t$. Assuming that $z_0 = z_0^1 = \ldots = z_0^{N-1} = \frac{\|y\|}{\|x\|}$, we see that those equals remain true throughout training. This gives us

$$z_t' = \frac{(z_t)^{N-1} \|x\|}{1 + e^{u(t)}}, \quad u(t) = (z_t)^N \|x\|, \quad u'(t) = N(z_t)^{N-1} z_t' \|x\|.$$  

Combining those equations gives us the ODE verified by the logit of our system

$$u'(t) = \frac{N \|x\|^2 \|z_t'\|^2}{1 + e^{u(t)}}.$$  

The solution of that equation for $N = 4$ and $N = 8$ can be found on Fig. 3 of the main text.

Appendix C.

C.1 Proof of Theorem 4.1

In this section we prove Theorem 4.1 from the main text on the dynamics of learning in the case of the Hinge loss:

$$L_H(W_t, Z_t; x) = \max\{0, 1 - Z_t^T (W_t x)_+ \cdot (1_{x \in D_1} - 1_{x \in D_2}) \}.$$  

Let us consider updates made after observing a point $x \in D_1$, the converse can be treated similarly with a simple change of sign. The system of ordinary differential equations verified by the parameters of our network is:

$$y_t' = \|x\|^2 z_t, \quad z_t' = y_t.$$  

The same relation between $y_t$ and $z_t$ appears in those equations than in the cross-entropy case. Defining $c$ similarly, we have $u'(t) = 2 \|x\| u(t)$ in the case where $c = 0$, leading to $u(t) = u_0 e^{\|x\|}$.

When $c \neq 0$, the same change of variables can be applied and leads to

$$y_t = \sqrt{c} \cosh(\frac{\theta_0}{2} + \|x\| t), \quad z_t = \sqrt{c} \sinh(\frac{\theta_0}{2} + \|x\| t), \quad u_t = \frac{c}{\|x\|} \sinh(\theta_0 + 2 \|x\| t),$$  

with $\theta_0 = \cosh^{-1}(\frac{\theta_0}{2} + \|x\|^2 z_0^2)$. Those equations are only valid until $u_t$ reaches 1.\footnote{We are considering class $\{1\}$ here, but the equivalent can be proven for class $\{-1\}$.} At that point learning stops, the network has converged. If the initialization is such that that condition is already verified, then the weights will not change as they already solve the task. The learning curves along with the initialization diagram can be found in Fig. 6. We notice a hard sigmoidal shape, corresponding to learning stopping when $u_t$ reaches 1.
C.2 General treatment

Let us now consider the general case of a class containing an arbitrary number of points $C_1 = \{x_i\}_{1 \leq i \leq m} \subset \mathbb{R}^d$. We consider the case of updates done in full batches (standard gradient descent in other words). In that case, assuming that one full pass on the dataset is done during each unit of time, we see that the network obeys the following dynamics:

$$w'_t = z_t \sum_{i=1}^m x_i^T, \quad z'_t = w_t \sum_{i=1}^m x_i^T.$$  

Letting $X = \sum_{i=1}^m x_i$ denote the sum of all the datapoints in that class, we see that those dynamics boil down to our previous treatment for a single point. The same cases appear, depending on the value of $c := \frac{1}{2}((w_0X)^2 - \|X\|_2^2 \alpha^2_{\delta})$. We explicitly treat the $c > 0$ case. Following the methods above, we see that: $w_t = y_t \frac{X^T}{\|X\|^2} + w_0^T$ where $y_t = \sqrt{e} \cosh(\frac{y_0}{2} + \|X\|t)$ and $w_0^T$ is the component of $w_0$ orthogonal to $X$ (and thus unchanged during training). With $z_t$ and $u_t$ defined as above (with $X$ instead of $x$), an arbitrary example $x$ is then classified as

$$P(x \in C_1) = z_t y_t \frac{X^T x}{\|X\|^2} + z_t w_0^T x = u_t \frac{X^T x}{\|X\|^2} + z_t w_0^T x.$$  

APPENDIX D.

In this section, we prove a relaxed version of Theorem 5.1 from the main text:

**Theorem D.2.** Let $\delta$ be our confidence requirement on class $D_1$ i.e. the training stops as soon as $\forall x \in D_1, P_t(x \in D_1) \geq 1 - \delta$. Let $t^*$ denote that instant i.e. $z_{t^*} \alpha_{t^*} = \log(\frac{1}{2})$. If $\beta_0 < 0$, the inequality (9) from the main text is valid. Otherwise, with $w_0 = (\alpha_0 x_1, \beta_0 x_2) + (x_1^+, x_2^+)$, we have

$$P_t((0, x_2) \in D_1) \leq \frac{1}{1 + e^{-\lambda \log(\frac{1}{2}) - z_{t^*} (\beta_0 - \alpha_0)}}.$$  

**Proof.** We start with the $\beta_0 < 0$ case. If $\beta_{t^*}$ is negative, the result from the main text clearly holds. Otherwise, there exists $t < t^*$ such that $\beta_t = 0$ ($\beta_t$ is increasing). The proof of Theorem 5.1 from the main text can then directly be applied to $[t, t^*]$. If $\beta_0 > 0$, the inequality on $\alpha_{t^*}$ and $\beta_{t^*}$ holds and gives $\beta_t \leq \beta_0 + (\alpha_t - \alpha_0) \lambda$. Plugging it into $P_t((0, x_2) \in D_1)$ concludes our proof.  

In the main text, we assume that $\beta_0 - \alpha_0 < 0$ and obtain a bound on the confidence which is independent from $\alpha_0$ and $\beta_0$. Using that bound allows to obtain Fig. 7 but is partly unfair as the initialization of the network is already favoring the strong feature.

However, we note that under small random initialization $z_{t^*}$ and $\alpha_{t^*}$ are of the same order of magnitude and $(\beta_0 - \alpha_0)$ is very small compared to $\log(\frac{1}{2})$. The additional term in the denominator thus has a limited effect on the exponential, gradient starvation is still happening (a fact confirmed by the experiment on the cats and dogs dataset). In the main text, Fig. 5 plots the upper bound for a fair initialization $\alpha_0 = \beta_0 = 0.1$ (in that case, we need to assume that $z_{t^*} = \alpha_{t^*}$).
APPENDIX E.

E.1 Mixture of Gaussian experiment

In this experiment, the data is constructed using eight independent Gaussian distributions around a unit circle. The variance of each Gaussian is chosen such that all eight modes of the data are separated by regions of low data probability, but still contain a reasonable amount of variance. This simple experiment resembles multi-modal datasets. Although this task might seem simple, in practice many generative adversarial networks fail to capture all the modes. This problem is generally known as mode collapse.

As shown in the main text, using the hinge loss instead of the common binary cross-entropy loss alleviates the problem significantly. The architectures used for the generator and discriminator both consist of four hidden layers where each layer has 256 hidden units. As a common choice, a ReLU is used as the non-linearity function for hidden units. The length of the noise input vector is 128. The Adam optimizer (Kingma & Ba, 2014) was applied during training with $\alpha = 10^{-4}$, $\beta_1 = 0.5$, and $\beta_2 = 0.9$. The PyTorch framework (Paszke et al., 2017) was used to conduct the experiment.

E.2 Dogs vs. Cats classification with light effect

For the purpose of highlighting the fact that the most frequent feature starved all the others, we conducted an experiment on a classification task. We modified the cats and dogs dataset (Kaggle, 2018) by setting the cats images to be lighter than the dogs images. To do so, each pixel in a cat image is scaled to be between 0 and 127 while each pixel in a dog image is scaled to be between 128 and 255. The dataset consists of 12500 images of each class. The classifier has an architecture similar to VGG16 (Simonyan & Zisserman, 2014). In order to isolate the effect of the induced bias, no regularization was applied. The Adam optimizer was applied here as well during training with $\alpha = 10^{-4}$, $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The PyTorch framework was used to conduct the experiment.