Topics in Machine Learning - Final Project

Marco Scialanga, Yerkin Yesbay

Abstract

We outline and discuss the main results of the paper *Gradient descent aligns the layers of deep linear networks* by Ji and Telgarsky[1] . We perform numerical experiments on the theorems proven in the paper. We make a few remarks (mostly absent from the analysis of the authors) on how the step size, initialization scale, width and depth of the networks affect the alignment speed. Then, we perform experiments with nonlinear networks and on nonlinearly separable data. Analyzing the results, we make a hypothesis regarding a possible alignment phenomenon and convergence to max-margin solution in these settings as well.

1 Introduction

Over the last few years, research in the theory of deep learning has made some progress in understanding how over-parametrized models trained via gradient descent obtain solutions with good generalization abilities. In certain simplified settings, it has been proven that gradient descent prefers low-complexity solutions even without explicit regularization (for example, Soudry et al. (2018)[2]). In the literature, this phenomenon is referred to as *implicit regularization* or *implicit bias*.

A common tool in the analysis of gradient descent is gradient flow, a continuous time version of gradient descent. Ji and Telgarsky (2018)[1] establish risk convergence and asymptotic weight matrix alignment for gradient flow applied to deep linear networks trained on linearly separable data, and similar results for gradient descent with a small enough (adaptively decreasing) learning rate. Furthermore, under mild additional assumptions on the training data set, a directional convergence to maximum margin solution is established. In this report, we will present the main results obtained by the authors, make some remarks on these theorems, and propose some hypotheses for different settings supported by numerical experiments.

2 Problem setting

We consider a binary classification problem with a data set $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$. Assume that the data is linearly separable, i.e. $\exists u \in \mathbb{R}^d$ such that $y_i \langle x_i, u\rangle > 0$ holds for any $1 \leq i \leq n$. Let $\bar{u} = \underset{\|u\|=1}{\arg\max} \underset{1\leq i\leq n}{\min} y_i \langle x_i, u\rangle$ denote the maximum margin solution with unit norm.

Consider a linear network of depth L parameterized by weight matrices $W_L, \ldots, W_1$, where $W_k \in \mathbb{R}^{d_k \times d_{k-1}}$, $d_0 = d$, and $d_L = 1$. Let $W = (W_L, \ldots, W_1)$ denote all parameters of the network. The (empirical) risk induced by the network is given by

$$\mathcal{R}(W) = \mathcal{R}(W_L, \ldots, W_1) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i W_L \cdots W_1 x_i\right) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(\langle w_{\text{prod}}, z_i\rangle\right),$$

where $w_{\text{prod}} := (W_L \cdots W_1)^\top$, and $z_i := y_i x_i$.

The paper considers gradient flow and gradient descent, where gradient flow $\{W(t) \mid t \geq 0, t \in \mathbb{R}\}$ can be interpreted as gradient descent with infinitesimal step sizes. It starts from some $W(0)$ at $t = 0$,

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

and proceeds as

$$\frac{\mathrm{d}W(t)}{\mathrm{d}t} = -\nabla\mathcal{R}(W(t)).$$

By contrast, gradient descent $\{W(t) \mid t = 0, 1, 2, \dots\}$ is a discrete-time process given by

$$W(t+1) = W(t) - \eta_t \nabla\mathcal{R}(W(t)),$$

where η_t is the step size at time t.

Before presenting the main results of the paper, we state the assumptions made by the authors. Assumptions 2.1 and 2.2 will be sufficient to prove results concerning gradient flow, while for theorems related to gradient descent it will be necessary to add assumptions 2.3 and 2.4.

Assumption 2.1 $\ell' < 0$ is continuous, $\lim_{x\to-\infty} \ell'(x) = \infty$ and $\lim_{x\to\infty} \ell(x) = 0$.

Assumption 2.2 The initialization $W(0)$ satisfies $\nabla\mathcal{R}(W(0)) \neq 0$ and $\mathcal{R}(W(0)) < \mathcal{R}(0) = l(0)$.

Assumption 2.3 ℓ' is β-Lipschitz and $|\ell'|$ is G-Lipschitz.

Assumption 2.4 For any $R > 0$, let $B(R)$ denote the ball $\{W | \max_{1\le k\le L} \|W_k\|\}$. Then, for every $R \geq 1$ consider a function $\beta(R) = 2L^2R^{2L-2}(\beta + G)$. Then, for gradient descent, assume the step size to be $\eta_t = \min\{\frac{1}{\beta(R_t)}, 1\}$, where R_t satisfies $W(t) \in B(R_t - 1)$, and if $W(t+1) \in B(R_t - 1),\ R_{t+1} = R_t$.

Assumption 2.5 The data points x_j for which $y_j\langle x_j, \bar{u}\rangle = \min_{1\le i\le n} y_i\langle x_i, \bar{u}\rangle$, i.e., the support vectors, span the whole $\mathbb{R}^d$.

We are now ready to present the focal contributions made by the authors.

3 Main results of the paper

Theorem 3.1 Under Assumptions 2.1 and 2.2, the below properties hold for gradient flow. If, in addition, Assumptions 2.3 and 2.4 are satisfied, then the theorem holds for gradient descent as well.

- $\lim_{t\to\infty} \mathcal{R}(W) = 0$.
- For any $1 \le k \le L$, $\lim_{t\to\infty} \|W_k\|_F = \infty$.
- For any $1 \le k \le L$, letting (u_k, v_k) denote the first left and right singular vectors of W_k,

$$\lim_{t\to\infty} \left\| \frac{W_k}{\|W_k\|_F} - u_k v_k^\top \right\|_F = 0. \tag{1}$$

  Moreover, for any $1 \le k < L$, $\lim_{t\to\infty} |\langle v_{k+1}, u_k\rangle| = 1$. As a result,

$$\lim_{t\to\infty} \left| \left\langle \frac{w_{\text{prod}}}{\prod_{k=1}^L \|W_k\|_F}, v_1 \right\rangle \right| = 1, \tag{2}$$

  and thus $\lim_{t\to\infty} \|w_{\text{prod}}\| = \infty$.

Furthermore, for exponential and logistic loss, one has the following convergence of w_{prod} in direction to maximum margin solution.

Theorem 3.2 Under Assumptions 2.2 and 2.5, for gradient flow $W(t)$ it holds that for almost all data and for losses $\ell_{\exp}$ and $\ell_{\log}$, then $\lim_{t\to\infty} |\langle v_1, \bar{u}\rangle| = 1$, where v_1 is the first right singular vector of W_1. As a result, $\lim_{t\to\infty} w_{\text{prod}} / \prod_{k=1}^L \|W_k\|_F = \bar{u}$.
If, in addition Assumption 2.4 is satisfied, the same holds for gradient descent.

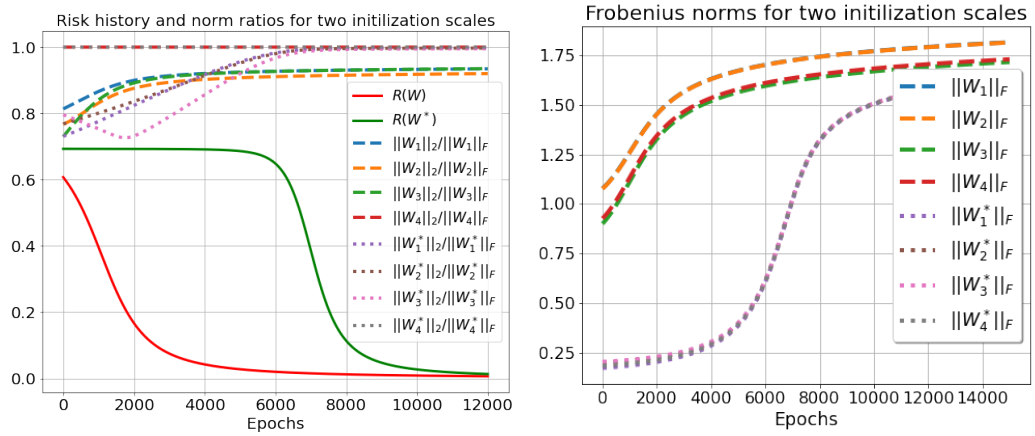

Figure 1: On the left: the empirical risk and norm ratios for two linear networks with weights $W = (W_1, W_2, W_3, W_4)$ and $W^* = (W_1^*, W_2^*, W_3^*, W_4^*)$. On the right: the Frobenius norms of the weight matrices. W^* has around 5 times smaller initialization scale than W.

Remark 3.3 Our experiments will mainly focus on the alignment phenomenon highlighted in 1.

4 Discussion and remarks on the alignment results

In this section we will present some remarks that we noticed when performing experiments on the theorems presented in the paper.

Remark 4.1 The learning rate suggested by the authors is quite small: while a constant step size of 10^{-2} is enough to minimize the risk, η_t can easily be needed to be $< 10^{-5}$ after a few thousands of iterations to reach full alignment, even for not so deep networks. This causes the layers' convergence to first rank approximations to be slow. However, at the same time, through numerical experiments, we saw that deviating from the suggested step size often resulted in the layers not approaching their normalized rank-1 approximations. This suggests that even if alignment phenomena were present in nonlinear deep networks (networks that are actually used), they would be difficult to achieve in practice, since the learning rates would have to be prohibitevly small, unless nonlinearity (or different optimizers) could allow for larger step sizes.

Remark 4.2 The width of the network itself is rather not so influential on the limiting behaviour, but the initial $\max_{1 \leq k \leq L} \|W_k\|_F$ matters a lot. Small enough initialization and the learning rate imposed by Assumption 2.5 ensure that $W_k / \|W_k\|_F$ will converge to 1 within a reasonable number of steps, according to our experiments. Figure 1 shows the evolution of risk, norm ratios and Frobenius norms of two linear networks with depth 4 and weight matrices 2×2 but different initialization scales. At some point the norm ratios of the layers of the network with bigger initial weights eventually almost flattens while still not being very close to 1.

Remark 4.3 For deeper networks it takes many more iterations to achieve alignment and small risk. The number of steps needed scales dramatically as the depth grows, as shown in Figure2. This is due to the definition of $\beta(R)$.

5 Numerical experiments

5.1 Experiments with ReLU and Variants

We now leave the scenario of the paper and turn our attention to non-linear networks and nonlinearly separable data. One of the simplest and most widely adopted activation functions is the Rectified Linear Unit (ReLU): $f(x) = \max(0, x)$. We are interested into how the phenomenon of the

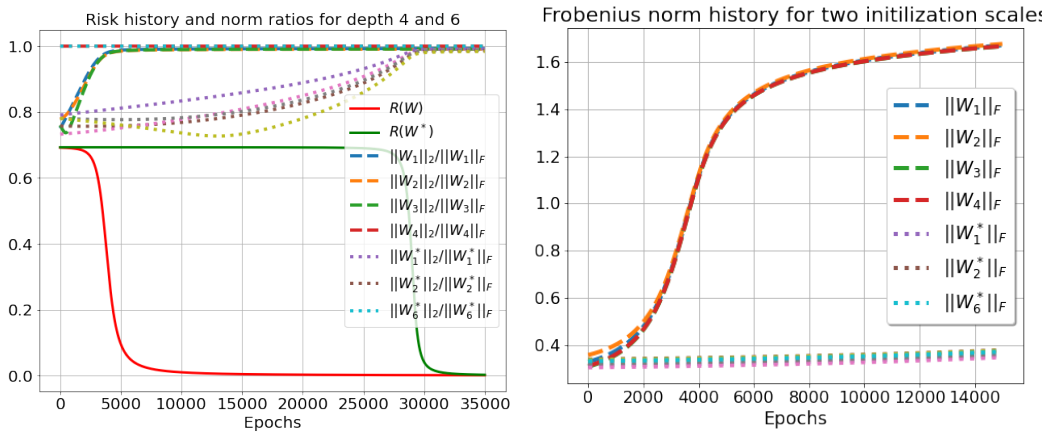

Figure 2: On the left: the risk and norm ratios for two linear networks of depth 4 and 6. On the right: their Frobenius norms.

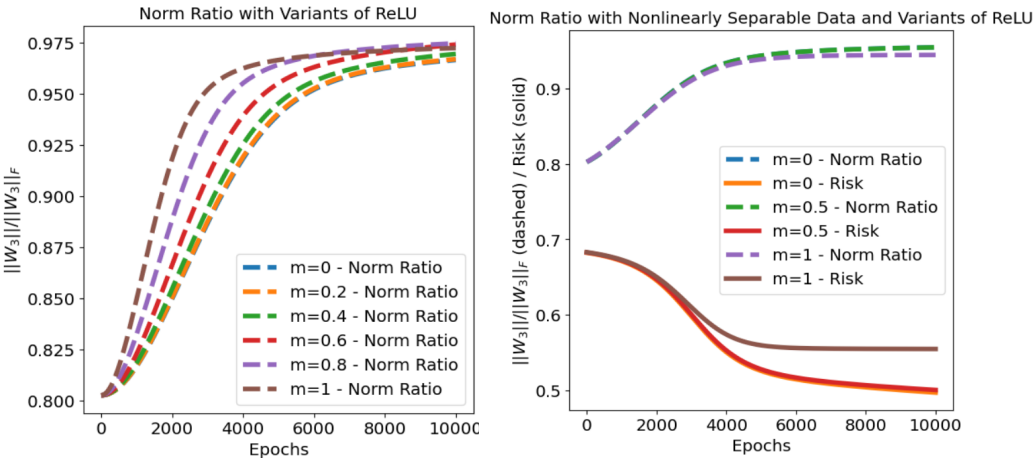

Figure 3: On the left: plot showing possible layer alignment for networks built with variants of ReLU. On the right: plot showing possible layer alignment even for linear network, to its rank-1 approximation despite data being only nonlinearly separable.

convergence of the normalized layers to their rank-1 approximations is affected by the introduction of nonlinearity in the model, for example with ReLU, or with variants such as:

$$f(x) = \begin{cases} mx & x < 0 \\ x & x \geq 0 \end{cases}, \ 0 \leq m \leq 1. \tag{3}$$

Thus, we repeated the same experiments as in Section 4, training a 4 layer neural network with 2×2 layers on a binary classification task with linearly separable data, but this time adding activation functions as in 3, trying $m = 0$ (ReLU), $0.2, 0.4, 0.6, 0.8, 1$ (identity). For step size, we use the same that was presented by the authors in the paper. As we can see in 3, it seems that we still obtain $\frac{||W_i||}{||W_i||_F} \to 1 \ \forall i = 1, \ldots, L$ (in the figure, we plotted such ratio for layer 3 only, but from the experiments we noticed that every layer had the same behaviour).
Furthermore, we decided to explore settings in which the data was nonlinearly separable. We generated our data with the `make_moons` function from the `sklearn` library (for reference: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html) and set the parameters to make sure that the data was separable, i.e., zero training loss could be achieved by a nonlinear neural network. As we can see from the plot on the right in 3, it seems as if the layers still converge to their normalized rank 1 approximation when trained with GD, even when the loss does not go to zero (as in the case of the linear network, i.e. $m = 1$).

In light of these experiments, we can make the following hypothesis:

Hypothesis 5.2 Under assumptions $2.1, 2.2, 2.3, 2.5$, logistic loss, nonlinearly separable data,

Remark 5.3 Another interesting aspect to investigate would be how m affects the rate of the alignment. 3 seems to suggests that there might be a relationship at play between these quantities.

5.2 Convergence to maximum margin when data is non-linearly separable

Consider a fully connected neural network of depth L with ReLU activation functions and let ℓ be the logistic loss. The logit output of the model is

$$f_W(x) = W_L \sigma_{L-1}(W_{L-1}\sigma_{L-2}(\dots(\sigma_1(W_1 x))),$$

where $\sigma_k : \mathbb{R}^{d_k} \to \mathbb{R}^{d_k}$ are entry-wise ReLUs. The first $L-1$ layers induce a feature map

$$\phi_{L-1}(x) = \sigma_{L-1}(W_{L-1}\sigma_{L-2}(\dots(\sigma_1(W_1 x))) \in \mathbb{R}^{d_{L-1}}$$

The last layer W_L can be viewed as a weight vector of a linear classifier in $\mathbb{R}^{d_{L-1}}$ trained on $(\phi_{L-1}(x_i), y_i)$. We are now interested in the following question: does this classifier tend to the maximum margin solution under gradient descent, provided that $\phi_{L-1}(x_i)$ are linearly separable after enough iterations? We trained such a neural network on nonlinearly separable data for enough iterations and trained a hard-margin linear SVM on the obtained $\phi_{L-1}(x_i)$. The maximum margin solution produced by SVD was closely aligned with W_L. Figure 4 illustrates the resulting decision boundaries of the two classifiers.

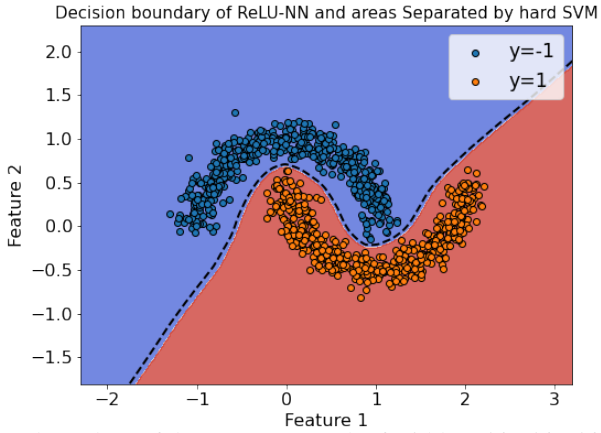

Figure 4: The decision boundary of the ReLU network of width (256,256,256) is shown by the dashed black line. The red and blue areas are separated by the decision boundary of a hard-margin SVM.

6 Conclusion

To conclude, the paper we examined shows a form of implicit regularization for GD applied to linear networks in a binary classification setting, with linearly separable data, for exponential and logistic loss. We made remarks on how the rate of convergence is affected by step size, initialization scale, width and depth of the networks. We also performed numerical experiments on binary classification with nonlinear networks trained with GD and nonlinearly separable data. These experiments gave interesting results and seem to suggest that alignment phenomena and convergence to max-margin solution might also be present in different settings than those presented in the paper.

References

[1] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

[2] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.