Gradient descent aligns the layers of deep linear networks

DongYoon Yang

Seoul National University

June 30, 2021

Table of Contents

1. Introduction
2. Notation, setting, assumptions
3. Results for gradient flow and gradient descent

Table of Contents

1 Introduction

2 Notation, setting, assumptions

3 Results for gradient flow and gradient descent

Introduction

- For deep linear neural networks on linearly separable data, gradient descent
  1. minimize the risk to 0.
  2. align adjacent weight matrices.(Rank-1 approximation)
  3. align the first right singular vector of the first layer to the maximum margin solution determined by data.
- By above alignment, neural network parameters in separable data converge to maximum margin solution.

Table of Contents

1 Introduction

2 Notation, setting, assumptions

3 Results for gradient flow and gradient descent

Notation

- $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$
- The data set is assumed to be linearly separable.
- $\gamma := \max_{\|u\|=1} \min_{1 \le i \le n} [y_i \langle u, x_i \rangle > 0]$: maximum margin.
- $\bar{u} := \operatorname{argmax}_{\|u\|=1} \min_{1 \le i \le n} [y_i \langle u, x_i \rangle > 0]$: maximum margin solution.
- $\boldsymbol{W} = (W_1, \cdots, W_L)$: all parameters of linear neural network. $W_k \in \mathbb{R}^{d_k \times d_{(k-1)}}, d_L = 1, d_0 = d$
- $W_{\text{prod}} := (W_L \cdots W_1) = \prod_{k=1}^{L} W_k$
- $\mathcal{R}(\boldsymbol{W}) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i W_L \cdots W_1 x_i)$

Assumption

Assumption 1.1.

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

Assumption 1.2.

The initialization $W(0)$ satisfies $\nabla \mathcal{R}(W(0)) \neq 0$ and $\mathcal{R}(W(0)) \leq \mathcal{R}(0) = \ell(0)$.

- Logistic loss and exponential loss satisfy assumption 1.1.
- W(0) is not critical point.
- Simply $W_1(0), W_L(0) \cdots W_2(0) \neq 0$ satisfy $\mathcal{R}(W(0)) \leq \mathcal{R}(0)$

Table of Contents

1 Introduction

2 Notation, setting, assumptions

3 Results for gradient flow and gradient descent

Gradient Flow, Gradient Descent

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

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

- Gradient flow can be interpreted as gradient descent with infinitesimal step sizes.

Risk convergence

Theorem

Under assumptions 1.1 and 1.2, the following properties hold:

- $\lim_{t\to\infty} \mathcal{R}(W) = \infty$
- *For any* $1 \le k \le L$, $\lim_{t\to\infty} \|W_k\| = \infty$
- *For any* $1 \le k \le L$,

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

and $\lim_{t\to\infty} |\langle v_{k+1}, u_k \rangle| = 1$

- $$\lim_{t\to\infty} \left| \left\langle \frac{W_{\mathrm{prod}}}{\prod_{k=1}^{L} \|W_k\|_F}, v_1 \right\rangle \right| = 1$$

and $\lim_{t\to\infty} \|W_{prod}\| = \infty$

Convergence to the maximum margin solution

- $\gamma := \max_{\|u\|=1} \min_{1 \le i \le n} [y_i \langle u, x_i \rangle > 0]$
- $\bar{u} := \operatorname{argmax}_{\|u\|=1} \min_{1 \le i \le n} [y_i \langle u, x_i \rangle > 0]$
- The data point z_i with $\langle \bar{u}, z_i \rangle = \gamma$ is called support vector.

Assumption 2.7.

The support vectors span the whole space $\mathbb{R}^d$

Assumption 3.5.

The step size $\eta_t = \min\{1/\beta(R_t)), 1\}$ where R_t satisfies $W(t) \in B(R_t - 1)$, and if $W(t+1) \in B(R_t - 1), R_{t+1} = R_t$.

- $B(R) = \{W : \|W\|_F \le R, 1 \le k \le L\}$
- $\beta(R) = 2L^2 R^{2L-2}(\beta + G)$

Convergence to the maximum margin solution

Theorem (Gradient Flow)

Under assumptions 1.2 and 2.7, for almost all data and $\ell_{exp}, \ell_{logistic}$

- $\lim_{t\to\infty} |\langle v_1, \bar{u}\rangle| = 1$
- $\lim_{t\to\infty} \frac{W_{prod}}{\prod_{k=1}^{L} \|W_k\|_F} = \bar{u}$

Corollary

Under assumptions 1.2 and 2.7, for almost all data and $\ell_{exp}, \ell_{logistic}$,

$$\min_i y_i \left(\frac{W_L}{\|W_L\|_F} \cdots \frac{W_1}{\|W_1\|_F}\right) x_i \xrightarrow[t\to\infty]{}$$

$$\max_{\substack{A_L \in \mathbb{R}^{1\times d_{L-1}} \\ \|A_L\|=1}} \cdots \max_{\substack{A_1 \in \mathbb{R}^{d_1 \times d} \\ \|A_1\|=1}} \min_i y_i \left(A_L \cdots A_1\right) x_i$$

Convergence to the maximum margin solution

Theorem (Gradient Descent)

Under assumptions 1.2, 2.7 and 3.5, for almost all data and $\ell_{exp}, \ell_{logistic}$

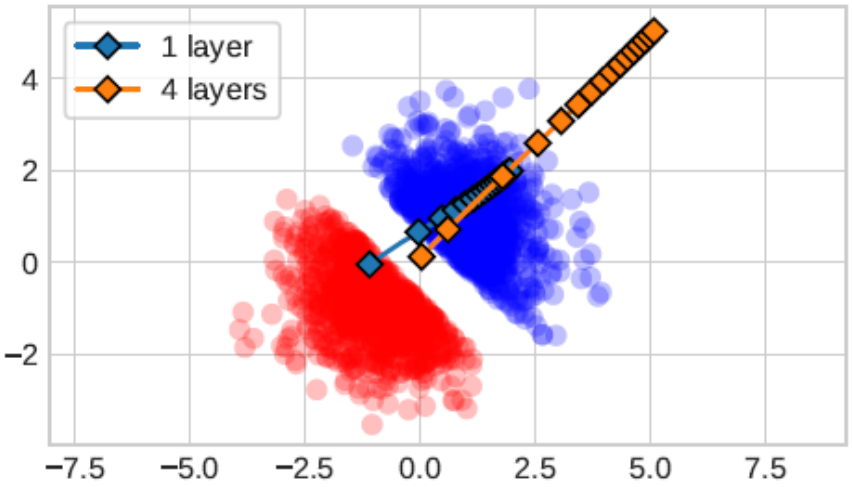
- $\lim_{t\to\infty} |\langle v_1(t), \bar{u}\rangle| = 1$
- $\lim_{t\to\infty} \frac{W_{prod}}{\prod_{k=1}^{L} \|W_k\|_F} = \bar{u}$

Corollary

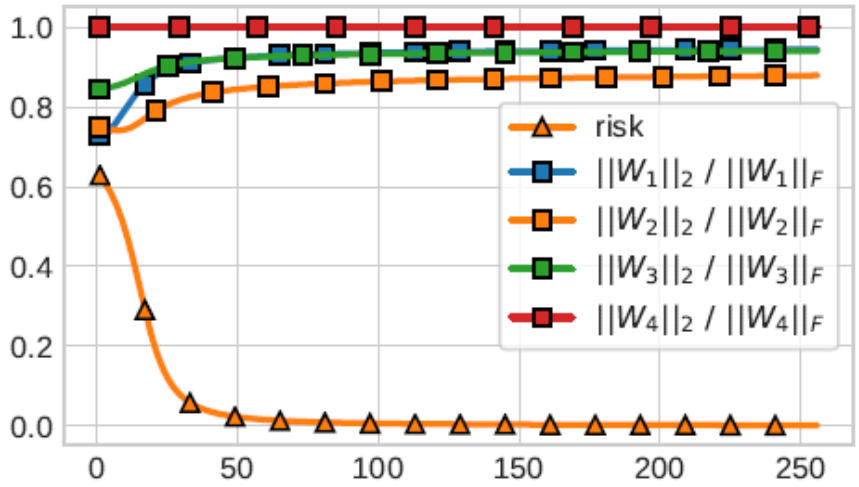
Under assumptions 1.2, 2.7 and 3.5, for almost all data and $\ell_{exp}, \ell_{logistic}$

$$\min_i y_i \left(\frac{W_L}{\|W_L\|_F} \cdots \frac{W_1}{\|W_1\|_F}\right) x_i \xrightarrow[t\to\infty]{}$$

$$\max_{\substack{A_L \in \mathbb{R}^{\mathbf{1}\times d_{L-1}} \\ \|A_L\|=1}} \cdots \max_{\substack{A_{\mathbf{1}} \in \mathbb{R}^{d_{\mathbf{1}}\times d} \\ \|A_{\mathbf{1}}\|=1}} \min_i y_i \left(A_L \cdots A_1\right) x_i$$



(a) Margin maximization.



(b) Alignment and risk minimization.

Figure 1: Visualization of margin maximization and self-regularization of layers on synthetic data with a 4-layer linear network compared to a 1-layer network (a linear predictor). Figure 1a shows the convergence of 1-layer and 4-layer networks to the same margin-maximizing linear predictor on positive (blue) and negative (red) separable data. Figure 1b shows the convergence of $\|W_i\|_2/\|W_i\|_F$ to 1 on each layer, plotted against the risk.

- In figure 1(a), the square indicates $W_{\text{prod}} = W_L \cdots W_1 \in \mathbb{R}^2$
- $\text{Rank}(A) = 1$ or $A = 0 \iff \|A\|_2 = \|A\|_F$

Summary

For deep linear neural networks on linearly separable data,

- $\frac{W_i}{\|W_i\|} \to u_i v_i^\top$.
- Adjacent rank-1 weight matrix approximations are aligned: $|v_{i+1}^\top u_i| \to 1$.
- The linear predictor induced by the net work, W_{prod} is aligned with the data: $\frac{W_{\text{prod}}}{\|W_{\text{prod}}\|} \to \bar{u}$.